

Holding the line at NASA

Space-based astronomy in the United States is under threat thanks to a misplaced sense of priorities within government. Researchers should take every opportunity to resist and to make the most of support from Congress.

For the past decade, a firewall has existed within NASA's budget, separating the agency's science programme from its astronaut programme. Thanks partly to congressional supervision, NASA has not raided its science account to pay for the space shuttle or the space station. Spending on science has actually gone up, while the shuttle and space-station budgets have stayed flat or declined.

Scientists believe, with good reason, that this is based largely on merit. The Hubble Space Telescope, the Wilkinson Microwave Anisotropy Probe and the Mars Exploration Rovers have returned solid results. The shuttle and space station have not.

Now, however, NASA faces more fiscal pressure than at any time in recent memory. Fixing the shuttle will, predictably, be more expensive than NASA first thought. The latest estimate is \$1.15 billion, plus the cost of added safety once the vehicle is flying again. Administrator Sean O'Keefe's decision to bar astronauts from visiting Hubble again has pinned the telescope's fate to an expensive robotic repair mission. Cost estimates vary wildly but start at about \$1.3 billion.

Looming even larger is President Bush's 'Vision for space exploration', which would send astronauts back to the Moon and on to Mars. No one is sure what this will cost, but the Congressional Budget Office recently guessed at \$127 billion by 2020. NASA wants to start next year, with \$438 million requested for the Crew Exploration Vehicle.

The White House and NASA sold the exploration programme on the basis that it would require no large increase in funding — the

money would come from retiring the shuttle and the space station. It is a sensible plan but may already be unravelling. A report released last week by the National Academy of Sciences showed how just one area of science — solar and space physics — could be affected by the "more constrained funding climate" that would accompany a Moon–Mars programme. A sequence of missions could delay high-priority projects such as the proposed Solar Probe for years.

Fortunately, Congress is watching. Last week the Senate appropriations committee cut NASA's \$16.2 billion request by \$664 million, including projects associated with the exploration programme. An additional \$800 million in emergency spending would go to solve the shuttle and Hubble problems. More importantly, the committee directed NASA not to upset the balance of its science programme to fund the Moon–Mars effort. The House appropriations committee, which made even deeper cuts to the exploration programme this summer, was more blunt, writing: "While the Committee is supportive of the exploration aspect of NASA's vision, the Committee does not believe it warrants top billing over science and aeronautics."

Bush's Vision has failed to wow either the media or the general public. NASA should stick to its original plan to pay for the Moon–Mars programme from within its existing spaceflight account. If the shuttle or space station have financial setbacks, the Vision should be scaled back or delayed. But hands off the science programme, which today is doing the real exploring in space. ■

Open-source biology

Researchers and entrepreneurs alike should welcome a move to develop a new commons in technological innovation.

With the research community increasingly frustrated by a growing forest of patents around innovations in the biological sciences, an initiative to make research tools from the life sciences open-source deserves to acquire some traction.

The Biological Innovation for Open Society (BIOS) initiative (see page 494) makes a distinction between tools and applications of innovation. Its champions argue that research tools should be freely available, much as operating systems, programming languages and standards are shared by the open-source software community.

The ranks of those wanting to see more put back into the life sciences for the public good are swelling. There is discontent that innovations from publicly funded scientists are being sold off to private companies and locked up in exclusive licensing. Concern is mounting that poorer nations are being further disenfranchised by richer countries' ownership and control of enabling technologies. Even purely academic scientists are not immune from the effect of licensing obstacles on new techniques.

BIOS should help on several fronts (see www.bios.net). Its intellectual-property database and associated informatics promise to bring more transparency to the opaque patent web and to provide tools to guide decision-making when choosing technologies. It will provide a suite of licences and other contract mechanisms for contributors to make their research tools available to a protective

commons. And it will provide an Internet-based mechanism to bring networks of researchers together to cooperate on specific technology development projects.

But are biologists ready for an open-source revolution similar to that which spurred information technology? The development of biological tools involves years of research and investment. And, in some cases, it's not as simple as downloading software, with potential logistical challenges involving the transfer of materials and issues of liability. But if BIOS can overcome these challenges, as its champions say it can, it offers a novel means for biologists to share research tools.

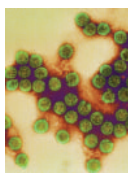
As its creators acknowledge, the success of BIOS hinges on active community participation — it needs a large network of scientists to devote their time to contributing research tools, sharing knowledge and addressing particular technology needs. It also needs financial support from private and public benefactors. If the open-source software movement is any guide, there will be willing backers of competitive open-source technologies — as exemplified by IBM's support of the Linux operating system.

The ultimate question is whether BIOS will benefit those for whom it is most intended and provide technology access to the poor and excluded. Whatever the outcome, it deserves a chance to garner support from the research community. At the very least, it may inspire other open-source initiatives within the biological sciences. ■





Dawdling probes
Funds sought to study slow speed of the Pioneers
p494



Plasmid pickle
Monkey virus may be exonerated of cancer link
p495



Broken bones
Corruption dogs sales of dinosaur fossils in China
p496



Seeing red
Russian Nobel laureate riled by US visa rules
p498

Global AIDS trial denied patients as US balks at generic drug use

Erika Check, Washington

US opposition to generic AIDS drugs is unnecessarily delaying the progress of a major international clinical trial, critics say.

The \$120-million trial, which is known as Strategies for Management of Anti-Retroviral Therapy, or SMART, is sponsored by the US National Institute of Allergy and Infectious Diseases (NIAID). It aims to enroll 6,000 patients to answer several questions about antiretroviral treatments for HIV, including whether they should be used early and aggressively, or whether treatment should only commence once a patient's immune system begins to fail. But the trial is running behind schedule because NIAID has balked at enrolling patients who take generic versions of the antiretroviral drugs.

According to trial organizers in the United States and elsewhere, patients in Argentina, Brazil and Peru were expected to join the trial earlier this year. But leaders of AIDS trials in South America and Thailand say they were told this summer that their patients would not be able to take part in SMART, because these people are all taking generic versions of brand-name AIDS medicines and those versions have not been approved by the US Food and Drug Administration (FDA). They say the instruction was handed down by the institute, which is part of the US health department's National Institutes of Health (NIH).

Written complaint

"We are told that, in SMART, no drugs not approved by FDA must be used," wrote Bernard Hirschel, leader of an AIDS trial covering Switzerland, Thailand and Australia, in an e-mail sent to colleagues earlier this month. A copy of his e-mail was given to *Nature* by a third party last week.

The Bush administration said earlier this



Brand rebuff: Thailand produces the world's cheapest anti-AIDS drugs.

year that it would not allow recipients of aid under the US President's Emergency Plan for AIDS Relief to buy generic drugs with aid money until they are approved by the FDA. But the NIH is not buying drugs for any patients in the SMART trial. Anthony Fauci, head of NIAID, says it is concerned about generic drugs only because they lack approval.

"NIH funded clinical trials in the developing world must meet the same standards as clinical trials conducted in the developed world, e.g. that the drugs being used are safe and meet accepted standards," Fauci wrote in a 16 September e-mail, in response to Hirschel's concerns.

But activists are sceptical about that

statement. "My sense is that what is driving this is a desire to protect the brand-name companies' market share," says Gregg Gonslaves, head of treatment and prevention advocacy at Gay Men's Health Crisis in New York.

Fauci told *Nature* on 23 September that the NIH had no blanket policy to exclude generic drugs from its international trials. He said the agency would evaluate countries on an individual basis.

"If a country has data that the drugs they're using appear to be comparable to the drugs in other countries, and they want to enroll in this trial, we see no reason why they have to switch to FDA-approved drugs," Fauci said. But he would not say whether any of the countries trying to get into the SMART trial would meet this standard.

Negotiations stalled

Meanwhile, the delays have kept hundreds of patients from joining SMART. Brazil, Argentina and Peru are waiting for NIAID to sign a contract confirming their participation, but that step has repeatedly been delayed, say researchers. The issue has also stalled negotiations with another important potential participant — Thailand.

The delays have implications for AIDS care around the world. SMART was expected to enroll 2,264 patients this year; but has so far enrolled only 792. The South American and Thai arms of the trial would add 733 patients to the total for this year.

"Peru, Brazil and Argentina are in a hold-ing pattern, and Thailand isn't even there because it got held up before it got started," says Fred Gordin, an infectious diseases specialist at the Veterans Affairs Medical Center in Washington DC and head of the group that has been funded by NIAID to run the SMART trial. "We really need these countries on board," he says.

Biologists launch 'open-source movement'

Carina Dennis, Sydney

An initiative is being established this week with a US\$1-million grant from the Rockefeller Foundation, to make research tools more readily available to biologists who could not otherwise afford them.

The Biological Innovation for Open Society (BIOS) initiative will seek to make information and technologies such as plant-breeding tools freely available. It will also provide scientists with better information about what they can access and, its founders hope, establish an international community of interested researchers.

Richard Jefferson is the initiative's leader and chairman of the Center for the Application for Molecular Biology to International Agriculture (CAMBIA), which is a non-profit research institute based in Canberra, Australia. He says BIOS could spur an "open source movement" in biotechnology, analogous to the one that has developed in the computer software industry.

Plant scientists in poor countries often complain that they are shut off from recent advances in agricultural biotechnology because they cannot afford licensing fees.

The initiative's first activities will be to gather a portfolio of research tools that can be used for free and to construct an easy-to-

use database of patent information. It will also provide templates of licensing agreements for scientists who want to make their technologies freely available. In turn, users will be obliged to freely release innovations based on these techniques.

Jefferson says that BIOS will encompass all forms of biological innovations, including agricultural and animal-breeding tools, genetic resources, medical treatments and environmental remedies. Its running costs will be covered by funds from sponsors and what he terms "non-compulsory" subscription fees paid by licensees.

Its initial portfolio of research tools will include a new method, developed by CAMBIA, for transferring genes into plants using modified bacterial species. Jefferson hopes to publish the technique shortly and says it will side-step patents held by biotech firm Monsanto on *Agrobacterium tumefaciens*, a bacterium currently used for this purpose. "It's a poster child for the initiative," says Jefferson of the new method.

Much of the initial, one-year Rockefeller grant is being spent on hiring patent and computer specialists to extend CAMBIA's patent database and draw up the licensing templates. IBM is also contributing computer hardware and software in order to

help get the initiative off the ground.

Some universities and companies already provide free licences for their technologies to researchers in poor countries. But Jefferson says these efforts are not sufficient to provide the "cooperative environment" that BIOS is setting out to build.

The success of the initiative will require the kind of community groundswell that buoys the open-source software movement, says Robert Zeigler of the Consultative Group on International Agricultural Research, a network of agricultural laboratories.

It will also require a cultural shift at universities, says Yochai Benkler, an information-law specialist at Yale University. "Many universities operate on the assumption that the role of their licensing policies is to maximize revenue," he says.

Although BIOS is expecting some resistance from biotechnology companies, a few of the larger companies have actually expressed support for it. "We have had discussions with BIOS and these will continue," says Ganesh Kishore, vice-president of technology at DuPont in St Louis, Missouri. "I don't view BIOS as a threat: it will be complementary. We need many innovations to build all the products that we want to build."

Support sought to investigate sluggish Pioneers

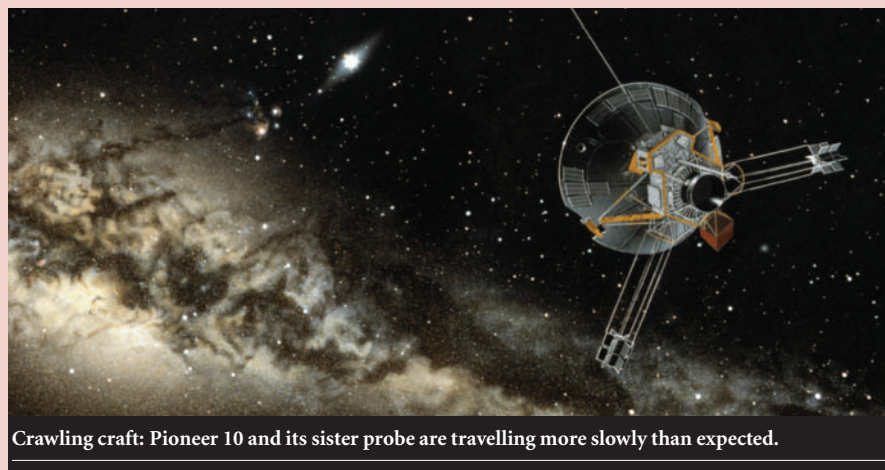
Jim Giles

Physicists are looking at ways to study a tiny but mysterious force acting on a pair of spacecraft at the edge of the Solar System.

One research team is seeking support to investigate flight data from the early stages of the Pioneer 10 and 11 missions, in a bid to understand why the craft are now travelling slightly more slowly than mission planners envisaged. A second, far more ambitious proposal would launch a dedicated space mission to study the effect, which has left the craft hundreds of kilometres closer to the Sun than had been anticipated.

The Pioneer anomaly, as the effect is known, was first detected about eight years after NASA launched the two craft to probe the outer Solar System in 1972 and 1973, respectively. It suggests that an unexpected force — about a hundred million times weaker than gravity's force on the surface of the Earth — has been pulling the craft back towards the Sun.

Many physicists think the slow-down is caused either by some unknown physical change in the craft themselves, increasing their drag, or by errors in the techniques used to track them.



Crawling craft: Pioneer 10 and its sister probe are travelling more slowly than expected.

But others have suggested that gravitational theory needs to be modified to account for the effect. This view was boosted in 2002, when an investigation into possible causes, led by John Anderson, a physicist at the Jet Propulsion Laboratory (JPL) in Pasadena, California, failed to find an alternative explanation (J. D. Anderson *et al. Phys. Rev. D* 65, 082004; 2002). The team rejected numerous effects, such as possible leaks of heat or gas from the crafts'

plutonium thermoelectric generators.

In the absence of a conventional explanation, physicists have produced speculative alternatives. Mordehai Milgrom of the Weizmann Institute of Science near Tel Aviv in Israel, for example, has developed a modified version of classical gravity in which the strength of the gravitational field decreases more slowly than newtonian theory suggests. Milgrom produced his theory to account for discrepancies in the

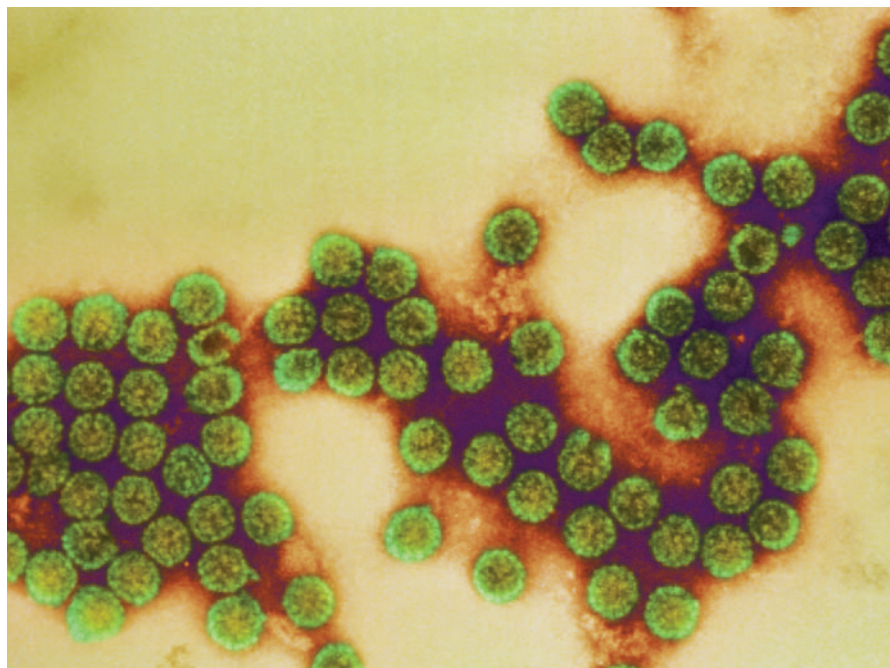
Monkey virus may be cleared of cancer link

Helen Pearson, New York

Laboratory contamination could have lent unwarranted support to the contentious idea that a monkey virus causes certain types of cancer, according to a study published last week.

The study tackled a long-standing disagreement in cancer biology about whether simian virus 40 (SV40), which contaminated stocks of polio vaccine in the 1950s and 1960s, could have infected vaccinated patients and triggered a range of chest, bone, brain and blood cancers. In support of this idea, some laboratories have identified fragments of the virus in tumour tissue samples — but others have struggled to repeat their results. The new study, led by pathologist Marc Ladanyi of the Memorial Sloan-Kettering Cancer Center, New York (F. López-Ríos, P. B. Illei, V. Rusch and M. Ladanyi *Lancet* **364**, 1157–1166; 2004), suggests a reason for this discrepancy.

“It’s a very major finding and may resolve the controversy,” says microbiologist Keerti Shah of the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, who has argued against a link between the virus and cancer. The study suggests that other researchers investigating the question may have inadvertently contaminated their



Simian scapegoat: some still argue that this monkey virus caused cancer through a contaminated vaccine.

motion of galaxies, an effect normally attributed to the presence of dark matter, but physicists have pointed out that it could also account for the Pioneer anomaly.

Anderson now wants to help resolve the uncertainty by reanalysing data from the first decade of the Pioneer missions. Slava Turyshev, a colleague of Anderson’s at JPL, estimates that it will cost about US\$250,000 to fund the analysis, and a grant application will be submitted to NASA later this year.

Turyshev and colleagues also have a grander plan: a dedicated spacecraft that would follow a similar trajectory to the Pioneer missions in a bid to recreate the anomaly. Their proposal, outlined in Paris on 16 September to an advisory panel of the European Space Agency (ESA), involves launching a spacecraft that would be followed a few kilometres behind by a reflective ball. Lasers on the craft would monitor the distance between the ball and craft, allowing researchers to detect and compensate for any acceleration caused by events on the craft, such as heat leaks.

But the craft would cost at least US\$500 million, and sources close to the panel suggest that it will not make the project one of the two priorities that it must convey to ESA next month. ■

tumour samples with plasmids — pieces of DNA that are widely used in molecular genetics laboratories and that frequently contain hidden fragments of SV40.

Ladanyi and his team searched for the virus in tissue samples of mesothelioma, a type of chest cancer. They used the polymerase chain reaction (PCR) to amplify a small stretch of DNA from the virus, and initially found that it was present in around 60% of their samples. But when they compared the genetic sequence of this region with others in gene databases, they realized that the section was also found in numerous laboratory plasmids.

Ladanyi and his team then carried out a series of detailed experiments showing that the virus DNA they had detected had come from contamination with plasmids, rather than the intact virus. Using genetic sequencing, for example, they showed that the virus fragment present in their samples contained a genetically engineered gap that is only found in plasmids.

Ladanyi believes other laboratories may have overlooked similar contamination in their PCR experiments, because they were unaware that so many plasmids contain sequences from SV40. To back up this argument, the team reanalysed the results of a 2002 study that identified the virus in brain and bone tumours (F. Martini *et al.* *Cancer* **94**, 1037–1048; 2002). The viral sequences this group had published showed that it too had detected plasmid sequences rather than SV40 itself.

The onus now falls on other laboratories to re-examine their results and rule out

contamination, says cancer epidemiologist Eric Engels, who studied the issue at the National Cancer Institute in Bethesda, Maryland. “Laboratories need to take this result seriously,” he says.

But others in the field who have detected the virus in their cancer samples stand by their original results. Virologist Janet Butel at the Baylor College of Medicine in Houston, Texas, says that her group has carefully checked its samples for plasmid contamination and can confidently rule it out. “This experiment does not mean all other laboratories are similarly contaminating their samples,” she says.

Those on both sides of the argument have other evidence to support their case. If the vaccine caused cancer, there should have been a rise in mesotheliomas or other cancers among those vaccinated decades ago with the virus-contaminated jab. But epidemiological studies have not detected one. On the other hand, hamsters and rats injected with SV40 go on to develop tumours.

Because the two sides are so entrenched, this single study is unlikely to resolve the debate, predicts Chris Wilson, an immunologist at the University of Washington in Seattle. Wilson served on a 2002 National Academies panel that assessed the evidence linking the virus and cancer. Either way, the study highlights the perils of PCR, which is notorious among researchers for producing false results in a variety of situations, even with the most experienced groups. “We know there are pitfalls,” says Wilson. “and one has to be extremely diligent to avoid them.” ■

Ancient ships lifted from Naples' railway tunnels

Federica Castellani

Italian builders have stumbled on the ancient Roman port of Neapolis while digging tunnels for an underground rail system in Naples. Archaeologists have now lifted three ships dating back to the first or second centuries AD to safety, and conservators are working on how to preserve this unusual treasure.

Workers found the first of the three wooden ships in January, buried near the Piazza del Municipio. The largest, which is 13.5 metres long and weighs some 25 tonnes, was removed by crane earlier this month. All are now stored in a climate-controlled facility in Piscinola.

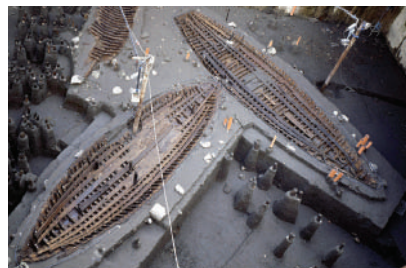
Researchers have long wanted to examine this ancient port. The harbour was abandoned at the end of the Roman empire, after which it gradually silted up; it had disappeared by the fourth century. This process created an airless environment which has kept the wooden ships in good condition, say researchers.

Historians expect an analysis of the wood to reveal something of the ancient climate and flora of the area. The cargo indicates how important Naples was for trade with the East. "Amazingly, the ships emerged with their entire cargo," says Daniela Giampaola, an archaeologist from Naples who is in charge of the excavation. Researchers have found ceramic containers of wine and cereals, corked glass perfume bottles and even sailors' shoe soles on the same site.

To protect the wood, conservators have encased the ships in fibreglass and filled these shells with distilled water. Chemists are testing the wood for damaging fungi and bacteria, and will then decide how best to treat it.

The ships will be kept in a room at a controlled temperature and oxygen level, under ultraviolet-filtered light.

Work on the tunnels is restarting, and researchers expect more treasures to appear as it continues. ■



Journey's end: these ships were entombed in mud by the fourth century.



Conflict is brewing in Liaoning, where famous fossils of bird-like dinosaurs have been found.

Feathers fly as China cracks down on illegal fossil sales

Rex Dalton, San Diego

Officials are taking a hard line on the illicit trade in dinosaur fossils in China, leaving palaeontologists worried that valuable specimens will be lost or destroyed.

This summer, fossil-rich quarries have been blasted, some 75 people threatened with arrest, and a potentially important new specimen apparently run over with a tractor — all in conflicts over fossils that could fetch a high price on the black market.

"It is unfortunate that the monetary value of these specimens places their existence and scientific study in jeopardy," says Mark Norell, chair of the Division of Paleontology at the American Museum of Natural History in New York City, who has worked in China. The only way to deal with the current situation is "to have laws with teeth that are enforced universally and equitably," he says.

In the northeastern province of Liaoning, fossils are typically dug up by poor farmers. Their sale is illegal — particularly for specimens of high scientific value. But a trade in run-of-the-mill fossils has been tolerated.

To ensure important fossils are available for study, some Chinese scientists have developed relationships with dealers and diggers. They buy the fossils or compensate the finder in other ways.

In June, as part of the crackdown, officials at the Liaoning Bureau of Land and Resources blasted 24 deep mine shafts that had been used by farmers to find fossils. Officials say this was done to stop the fossils from being stolen. But scientists contacted by *Nature*, who

requested anonymity for fear of jeopardizing any future work in Liaoning, describe the move as "sabotage", and say it has shut down sites that were producing quality specimens.

This summer, police scrutinized about 75 dealers and farmers; some were held briefly, charged nominal fines and then released. At the same time, informed scientists say, the provincial government has taken no action against major fossil markets, including a substantial one in the town of Chaoyang that functions openly. The scientists say that the action against the farmers may have been an attempt at extortion.

The conflict reached the attention of international dealers when a potentially exciting fossil was caught up in it. This fossil, which may represent an undescribed bird-like dinosaur, was found in July at the famous site of Sihetun during the building of a museum. Some reports say that the specimen included near-complete fossilized remnants of feathers, pressed into the rock along with the 2-metre-long skeleton after a volcanic explosion some 120 million years ago. Feather material from the Cretaceous period would be scientifically important — and would make the specimen very valuable.

Farmers are reported to have crushed the fossil with a tractor rather than give it up to officials they believed to be corrupt. Pieces are being smuggled away from the site, one source said. Chinese scientists were unable to provide more detail about the incident, but agreed that the location of a potentially exciting new specimen is now in doubt. ■

Beagle cash dogged by dissent over wording

Mark Peplow, London

Britain's failed Mars lander, Beagle 2, has ignited a row over money — €16 million (US\$20 million) that the European Space Agency (ESA) provided four years ago to keep the project going.

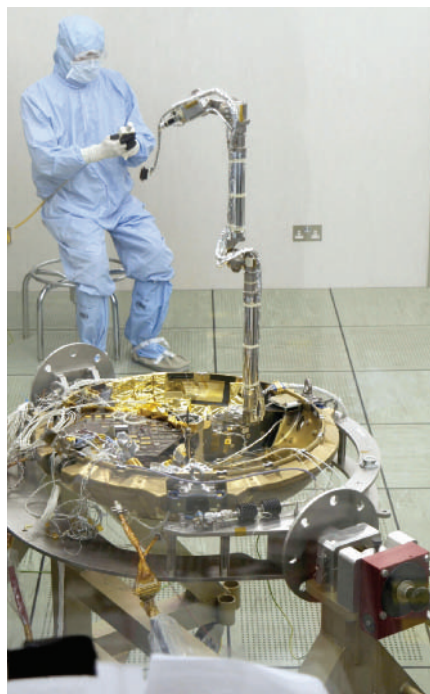
Disagreement about the conditions under which the money was handed over — and whether Britain is to repay it — has surfaced at senior levels in the space agency. David Southwood, ESA's scientific director, says it is unlikely that the agency will again allocate so much cash to a project on such an informal basis.

The transfer was agreed by ESA's Science Programme Committee, which allocates funds for agency projects, at two meetings in October and November 2000. The committee decided that ESA should allot €24 million to the Beagle probe, a late addition to the agency's Mars mission. The probe subsequently went missing while attempting to land on Mars on 19 December 2003.

One-third of the money went to preparing ESA's Mars Express spacecraft to carry the probe. But two-thirds of it was described at the meetings as a "loan", senior scientists tell *Nature*.

"The word 'loan' was used," said Risto Pellinen in an interview. Pellinen, an atmospheric physicist at the Finnish Meteorological Institute in Helsinki, was on the committee at the time and now chairs it. But he says the money was not in fact a loan, and was not meant to be paid back as cash. The word was used, he says, "because we are not native English speakers — we can't fine-tune our language".

The minutes of the October meeting state



Funding for Beagle 2, which was later lost on Mars, is still regarded by some as a 'loan'.

that "the science programme would be reimbursed the sum of €16 million" by Britain. The committee members, and Southwood, say that the wording was meant to imply that Britain would repay the contribution by providing goods and services to that value for future ESA missions.

The meetings discussed the possibility of a UK contribution towards Gaia, a proposed star-mapping mission due for launch some time after 2010. But four years after the

agreement, there is no indication of what Britain is planning to do. "What leaves a funny taste is this attitude that no one speaks about it any more," complains one senior space scientist who is knowledgeable about the arrangement but declined to be identified. He suggests that everyone would be happier if Britain repaid the money in cash: "A cheque for €16 million would be best."

Southwood says that David Sainsbury, the UK science minister, wrote to the committee at the time promising to do his best to recompense ESA. But Southwood insists that no one expected it to be paid back in cash: "It was a gentleman's agreement. You can't really call it a loan, it was a *quid pro quo*."

He concedes that it is unlikely that ESA will ever again grant so much money on such a basis. "I'm not convinced people are gentlemen any more," he adds ruefully.

Britain has no firm plan for how it will reimburse the space agency, says David Leadbeater, deputy director-general of the British National Space Centre. "There's nothing that says this must be sorted out in the context of a particular mission," he says, adding that he expects that the money will be repaid through contributions in kind to future projects. "There is still a commitment to resolve this to the satisfaction of all ESA members," insists Leadbeater.

Pellinen defends the arrangement. "In my opinion, this was the only way to get things done in the available time," he says. But he thinks negotiations about recompense are pointless until Gaia is ready for development in late 2005 — after the end of his chairmanship. "This matter will not be forgotten," he promises. ■

NIH researchers face blanket consulting ban

Emma Marris, Washington

The National Institutes of Health (NIH) has announced plans to ban its scientists from carrying out any paid consultancy work for biotechnology and pharmaceutical companies for 12 months.

In a memo to staff on 24 September, NIH deputy director Raynard Kington said that an internal ethics investigation had "identified vulnerabilities in our system that give us pause".

The ban is a blow to the large scientific staff on the biomedical research agency's main campus in Bethesda, Maryland. "I think it's pretty demoralizing," said one staff member, who declined to be identified. "We're being held to a very different standard from our colleagues elsewhere."

In July, the Office of Government Ethics

(OGE) issued a sharp review of the NIH's ethics rules and its plans to modify them. The OGE is now being asked to approve the proposed moratorium. Once it has done so, a start date for the ban will be set, says Kington. That could take several months.

Kington acknowledges that it will be difficult to recruit top scientists to work at the NIH if they can't supplement their government salaries. "We will do as much as we can to make sure we have the best and the brightest," he says, although he admits that "this may very well have a negative effect".

The NIH has been under intense congressional and public scrutiny since a story in the *Los Angeles Times* last December revealed consulting arrangements and lecture awards that, the paper said, posed conflicts of interest (see *Nature* 426, 741; 2003).

The OGE blamed the problem on what it termed a "permissive climate" for deals between scientists and industry. This was implemented in 1995 by the NIH director at the time, Harold Varmus, in a bid to attract scientific talent to the agency's research laboratories from the universities.

Varmus, who now supports a ban on consulting for senior staff, says he hopes that the wider ban will not become permanent. "A complete long-term ban would be a disaster for the intramural programme," he says. "The NIH became a very attractive venue when we made it more like the outside world." ■

For more news and analysis go to
news@nature.com
www.nature.com/news

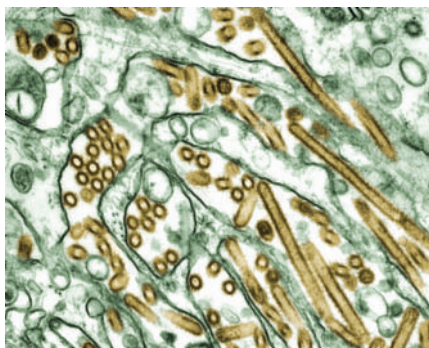
US orders bird flu vaccine before start of clinical trials

Washington The US Department of Health and Human Services last week ordered 2 million doses of vaccine to fight bird flu — even though the jab has yet to enter clinical trials.

The prototype vaccine, to be made by Aventis Pasteur, is designed to fight the H5N1 strain of avian influenza, which has devastated poultry flocks in Asia, killed nearly 30 people and sparked fears of a human pandemic. Small-scale clinical trials of the vaccine are planned for before early 2005 to determine its safety and dose.

Bruce Gellin, director of the department's National Vaccine Program Office in Washington, says the vaccine order is a trial run designed to reveal any obstacles in large-scale manufacture, such as whether there are problems growing the virus in hens' eggs.

Gellin says the vaccine is likely to sit on the shelf until the results of the clinical trials are in. The vaccine may even be thrown away, he acknowledges, if it proves unsafe or if the bird-flu virus mutates, rendering the vaccine ineffective.



Cause for concern: the H5N1 strain of bird flu virus (gold) sparked fears of a human pandemic.

Fossils figure in multiple guises, researcher claims

London Imitation may be a form of flattery, but it tends not to be appreciated when it comes to scientific papers.

Such is the problem faced by Mostafa Imam of the College of Education for Girls in Madinah Al Munawara, Saudi Arabia. Allegations have surfaced that Imam included in his papers photographs of bacterial fossils taken by fellow palaeontologists, with captions indicating that the fossils were found in other sites and showed different species.

Julio Aguirre, a palaeontologist at the University of Granada in Spain, investigated Imam's work after being sent one of his

Russian Nobel laureate left out in the cold

Moscow A Nobel Prize-winner who had planned to leave his native Russia to deliver a lecture in the United States last week instead stormed out of the US consulate in St Petersburg without a visa.

Zhores Alferov (right), who received the 2000 Nobel Prize in Physics for his work on semi-conductors, and now directs the Ioffe Physico-Technical Institute in St Petersburg, had been invited to lecture at the University of California, Berkeley, by the university's regents.

But when trying to get a visa, Alferov reportedly became indignant at being repeatedly asked by a consular official about the nature of



his work. When it became clear that the visa would not automatically be granted, Alferov left the building.

A spokesman for the university said it hoped to reschedule the talk for next spring.

papers to review. He claimed last month that Imam has produced a string of publications by lifting photographs from other papers (J. Aguirre *Rev. Esp. Micropaleontol.* 36, 349–352; 2004). Aguirre is calling on palaeontology journals to check Imam's publications for plagiarism.

Imam says that many of Aguirre's accusations are "lies". He says one of his papers was printed without his having seen the proofs, and that some of the text and pictures had somehow been changed in the process.

France gives research a billion-euro boost

Paris The French government announced an increase in the civil research budget for 2005 of €1 billion (US\$1.2 billion) last week — a rise of 10% from 2004.

In so doing it kept the promise it made in March to a research community angered by job cuts and budget freezes (see *Nature* 428, 105; 2004). But with no new jobs being created in research-agency labs by the funding, the reaction from many researchers was negative.

A third of the money will be used to create a new national research agency; another third will promote development in industry through tax breaks; and the final third will go to public research laboratories.

Presenting the budget last Wednesday, research minister François d'Aubert said that the budget increase was only one step towards necessary reform. A forthcoming white paper on science, which will be debated later this year, is expected to lead to a long-awaited overhaul of the French research system.

Smithsonian set to launch a whale of an exhibit

Washington The Smithsonian's National Museum of Natural History is planning to embark on a \$60-million ocean initiative. It will fund research and a museum exhibit on ocean sciences that is due to open in 2008.

The exhibit will be dominated by a full-size model of a northern right whale, which will hang from the ceiling in the main hall. Multimedia presentations will tackle climate change and other hot topics, and will be updated as the science changes. "If the hall were up right now, the hurricanes Ivan and Jeanne would be being tracked in real time," says Robert Sullivan, associate director for public programmes at the museum.

The National Oceanic and Atmospheric Administration has provided \$18 million in funding, which the Smithsonian has matched.

Science committee chairman has heart op

Washington Congressman Sherwood Boehlert (Republican, New York), chairman of the US House science committee, underwent coronary artery bypass surgery on Monday this week.

Abnormalities in the congressman's heart were found during a routine physical earlier this month. He checked into the National Naval Medical Centre in Bethesda, Maryland, where doctors discovered blockages in several arteries.

Boehlert, who was 68 on Tuesday, is up for re-election in November. Joe Pouliot, a spokesman for the science committee, says the surgery will delay some hearings, but expects the committee to pick up where it left off after the election.

A dangerous elixir?

Testosterone therapy jacks up vigour, sex drive and mental acuity — or so proponents claim. But are those who experiment with this potent sex hormone gambling with their health? Helen Pearson investigates.

It injects a spring into the step, a lift to the libido and a boost to the brain — enthusiasts for testosterone replacement therapy make it sound like a smart, super-charged version of Viagra. No wonder, then, that growing numbers of men — and even a few women — are dosing themselves with the male sex hormone.

In 2003, according to the pharmaceutical consulting company IMS Health, based in Fairfield, Connecticut, the number of American men being prescribed testosterone was over 2 million, having more than doubled from around 900,000 in 1999. The figure is thought to be rising still.

Most of these men are concerned about losing their youthful vigour. But in the medical world, controversy is raging over whether otherwise healthy men, whose testosterone wanes naturally with age, are likely to derive any benefit from such treatments. From what we know about the hormone, testosterone replacement might also accelerate the onset of prostate cancer. All the more worrying, then, that testosterone therapy is likely to be lifelong for some men: almost 30% of those using the hormone are aged between 18 and 45.

The big problem is a dearth of clinical data from which to assess risks and benefits. Indeed, experts say that patients taking testosterone are literally experimenting with their health. "The momentum towards testosterone replacement is reaching a point where it needs serious study to see if it should be supported by the scientific community, or set aside," says Bill Hazzard, a geriatrician at the University of Washington School of Medicine in Seattle.

On trial

Hazzard was a member of an expert panel convened by the US Institute of Medicine that last year recommended setting up a raft of short-term trials to better establish the effect of testosterone therapy¹. The National Institute on Aging in Bethesda, Maryland, says that it will start organizing such trials in elderly men — who are most likely to benefit from replacement therapy — later this year.

But these trials are unlikely to end the controversy. Some researchers argue that only large, long-term studies will reveal the subtle benefits and risks of testosterone therapy. "I think it's almost unethical not to do a study like this," says gerontologist Alvin Matsumoto, also at the University of Washington.

Talk to someone who has taken testosterone, and it's easy to understand why the



Promise of youth: gels that can be simply rubbed on the skin have boosted testosterone's use.

hormone is in hot demand. Joe Marcklinger, a 58-year-old land surveyor from Boston, turned to testosterone two years ago when he was suffering from depression. After rubbing on daily doses of hormone-laced gel, he not only threw out his antidepressants but had more energy and muscle tone, and a healthy appetite for food and sex. "I felt pretty darn good," he says. "It turned the clock back ten years."

In addition to its effects on the brain and behaviour, testosterone fuels the manufacture of men's sperm, promotes muscle growth and strengthens bones. On the downside, it contributes to male pattern baldness.

Testosterone production by the testes surges upwards around puberty and remains high throughout a man's twenties and thirties. Its levels then dwindle by about 1% a year². This natural decline has been dubbed the 'andropause', because of its parallels with the female menopause. Women also produce testosterone — albeit at less than 10% of the levels found in men — from their ovaries and adrenal glands. Again, production of the hormone seems to decline with age.

The case for testosterone therapy in men

stems from experience with hypogonadism, in which hormone production from the testes is extremely low. Young men with the condition suffer from symptoms associated with ageing — including a loss of muscle, sex drive and mental acuity, plus a gain in flab. Testosterone is an effective treatment.

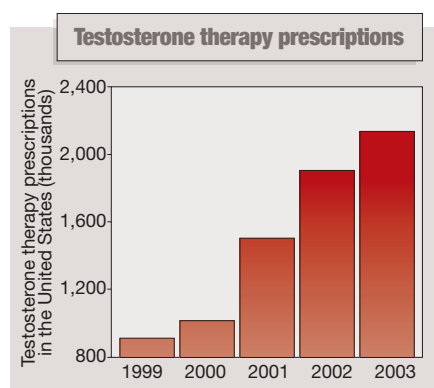
Unreliable evidence

Doctors had wondered for a long time whether ageing men, whose hormones flag to similar levels, might also gain a pick-me-up. But the use of testosterone really took off after the mid-1990s, when injections were replaced with easy-to-use gels and patches. The hormone can also shut down sperm production, which is why it is being investigated as a male contraceptive.

One of the fiercest debates among medical specialists is whether the natural decline in testosterone production has any ill effects. Some experts point to an array of studies linking the decline to failing strength, libido and bone density. But others argue that these naturally deteriorate with age, making it difficult to single out testosterone as a cause. "It might play a role but it's certainly not the



Some men hope that testosterone therapy will restore their strength and vitality.



only answer," says Lisa Tenover, a geriatrician at Emory University in Atlanta, Georgia.

The second controversy is whether otherwise fit middle-aged and older men benefit from increased testosterone. In one frequently cited trial, led by Peter Snyder at the University of Pennsylvania in Philadelphia, men over 65 who wore a testosterone patch for 36 months gained more muscle and lost more fat than those who donned a placebo³. But in its 2003 report, the Institute of Medicine found that the vast majority of the 31

trials in the literature were small, short and inconclusive.

Then there is the concern that testosterone replacement might increase the likelihood that latent cancerous cells in the prostate gland will transform into tumours. Some cancer therapies slow the growth of prostate tumours by reducing levels of testosterone, but there are few data linking testosterone therapy to an elevated cancer risk⁴. "There's still no convincing evidence," says Darracott Vaughan, a urologist at Cornell University's Weill Medical College in New York, and a member of the Institute of Medicine panel.

The uncertainty is compounded by the unreliability of the techniques commonly used to measure levels of testosterone in the blood. An evaluation earlier this year, led by Christina Wang of the University of California, Los Angeles, found that eight assays that used antibodies to gauge testosterone in the blood often produced over- or under-estimates compared with the gold-standard method of chromatography⁵.

Given these questions, Matsumoto and Glen Cunningham of Baylor College of Medicine in Houston, Texas, teamed up four years ago to propose a large clinical trial to study testosterone's effects. The trial, which grew after revisions into a \$120-million, six-year, project across 40 sites involving 6,000 men, gained support from the National Institute on Aging, the Department of Veterans Affairs and the drug industry.

Cancer fears

But the plan ran into trouble in 2002, when the consortium went looking for financial backing from other institutes at the National Institutes of Health's Bethesda campus. At the time, concerns about hormone therapy in general were running high because of the Women's Health Initiative, a vast study of female hormone replacement that was halted in 2002 after the benefits proved to be outweighed by long-term risks of heart disease, breast cancer and stroke⁶. The halting of that trial led researchers to advise that women on long-term courses of oestrogen and progestin re-evaluate their treatment.

Andrew von Eschenbach, director of the National Cancer Institute, who had previously headed a prostate-cancer research programme, weighed in against the study, citing fears that testosterone could encourage prostate cancer in some of the trial participants. His objection, among other factors, ensured that the trial remained stalled on the starting blocks. "We were disappointed, of course," says Cunningham.

Called in to map a way forward, the Institute of Medicine panel pinpointed the need for more evidence of testosterone's benefits

before embarking on a large-scale trial. Its report recommended a series of smaller investigations looking for gains in strength, sexual function, cognition and quality of life in men aged 65 and over.

But Matsumoto, Cunningham and others claim that the small trials will struggle to show subtle benefits in only a year or two. Far better, they say, to launch a large trial that could — like the Women's Health Initiative — be halted if it revealed serious risks. "They did a major disservice by not suggesting a large trial," argues John Morley of St Louis University School of Medicine, Missouri, a proponent of testosterone therapy. "By the time these studies come out, the baby boomers will have taken testosterone for ten years."

While the debate over male testosterone replacement rumbles on, women are getting in on the act. Most of the attention so far has focused on the hormone's use for treating female sexual dysfunction. One trial, led by Jan Shifren of Massachusetts General

Hospital in Boston, showed that women suffering from low testosterone levels after having their ovaries removed had sex more often after wearing a testosterone patch than those given a placebo⁶. Proctor & Gamble is now seeking regulatory approval to begin market-

ing a testosterone patch designed for use in women in 2005; other pharmaceutical companies are developing similar products.

Once such testosterone formulas hit the market, they raise the prospect that some women will use them as a 'lifestyle' drug — just as Viagra is used by some men who don't have major problems with their potency. Such a prospect worries endocrinologists, who point out that the hormone may prompt acne, the growth of body hair — and perhaps unknown health problems if it is taken for long periods without proper monitoring.

Given this, some experts are calling for larger and longer trials on the safety and efficacy of testosterone therapy in women. "Otherwise, there'll be a big uncontrolled human experiment," says Susan Davis, an endocrinologist at Monash University near Melbourne in Australia. In men, of course, the equivalent experiment is already up and running.

Helen Pearson is a reporter for news@nature.com and is based in New York.

1. Liverman, C. T. & Blazer, D. G. (eds) *Testosterone and Aging: Clinical Research Directions* (National Academies, Washington DC, 2003).
2. Gray, A., Feldman, H. A., McKinlay, J. B. & Longcope, C. J. *Clin. Endocrinol. Metab.* **73**, 1016–1025 (1991).
3. Snyder, P. J. et al. *J. Clin. Endocrinol. Metab.* **84**, 2647–2653 (1999).
4. Rhoden, E. L. & Morgentaler, A. *N. Engl. J. Med.* **350**, 482–492 (2004).
5. Wang, C., Catlin, D. H., Demers, L. M., Starcevic, B. & Swerdloff, R. S. *J. Clin. Endocrinol. Metab.* **89**, 534–543 (2004).
6. Rossouw, J. E. et al. *J. Am. Med. Assoc.* **288**, 321–333 (2002).

Fishing for trouble

Plans to push tuna farms out into open waters off the coast of the United States are raising an environmental alarm.

Rex Dalton discovers the kind of problems these offshore ranches might cause.

In the blue waters off the coast of Baja California, Mexico, circular nets are buffeted by the teeming mass of tuna inside. Every day boats pull up to these ocean 'ranches' and workers toss sardines into the churning water, fattening their investments while keeping an eye out for predatory sea lions.

The bluefin tuna being tended in these pens are destined for the sushi markets of Japan, which last year paid Mexican ranchers about US\$50 million for the prized fish. Such ranches are turning into big business around the world. Some countries in the Mediterranean farm the fish. And off the coast of southern Australia, more than 100 pens operated by a dozen firms are harvesting even bigger sales. Now, in the United States, legislation is quietly being drawn up to facilitate such fish-farming operations in offshore waters — beyond the environmental control of coastal states, in waters difficult for anyone to police.

This expansion has some people worried, as the history of tuna farms from Mexico to Australia shows that these ranches can cause damage to the marine environment. Coastal residents near Ensenada, Mexico, have complained in the past that some ranch operators have shot sea lions to protect their fish. And scientists are still trying to determine whether fish-food imported to Australian tuna farms was the source of a virus that wiped out the sardine fishery along that continent's southern coast in the late 1990s.

Seabirds were starved by that disaster, and fishermen left idle.

Some fear similar losses if ranches start to sprout in deep waters far from the US coast. "The opportunity for large-scale environmental disasters is enormous," says John Volpe, a fisheries ecologist at the University of Alberta in Edmonton. Volpe was one of the first scientists to issue warnings about potential environmental damage from salmon pens off British Columbia in the mid-1990s. Such farms are now blamed for spreading disease and lice into wild populations, and polluting the local environment. "We are one season from having wild salmon wiped out by lice," says Volpe. Now that similar warnings are being sounded for offshore tuna farms, he hopes someone will listen.

Offshore investments

But administrators at the US National Marine Fisheries Service (NMFS), an agency within the Department of Commerce's National Oceanic and Atmospheric Administration, are enthusiastically backing the concept. NMFS officials have been drawing up legislation and formulating plans for offshore aquaculture for years, as a way to provide more home-grown fish and aid the local economy.

In recent months, the NMFS has circulated draft legislation to federal agencies for comment. The bill is designed to permit aquaculture in waters outside the 5.5-km boundary of state control, but within the 370-km Exclusive

Economic Zone of the United States.

However, details of the proposal are tightly guarded: even the NMFS's own Marine Fisheries Advisory Committee (MAFAC) was given only a verbal summary of the legislation when it was asked to provide guidance on the concept in August. Environmental organizations have been trying to provide input for more than a year, but say they have been given the brush-off. The draft legislation is reportedly now close to being presented to Congress, and environmentalists fear it will slip through in the final weeks before the election.

With legislation pending, a research-business consortium is making plans for a tuna ranch in the Santa Barbara Channel off California. The consortium, led by the Hubbs-SeaWorld Research Institute of San Diego, hopes to anchor two square kilometres of nets on a former Chevron oil-drilling platform, about 20 km off the coast, and fill the nets with tuna and other deep-water fish.

The project would begin as a research facility, examining the capability of offshore farms and their environmental impact. The non-profit institute says it also has plans to extend operations to a commercial venture, using millions of dollars from fish sales to support the facility and its research. Chevron is funding the institute's start-up costs, and offering \$10 million to run the operation for three years. The oil company hopes to avoid the substantial expense of removing the oil platform completely.



B. BARBEY/MAGNUM

Down on the farm: critics condemn the ranching of bluefin tuna, which goes to Japan's markets (above).

California, like other US states, only controls waters up to 5.5 km from the shore, so the consortium has only needed to apply for federal permits. But environmentalists and officials in California are worried about the plans and threaten legal action if the state gets no say on the project. "They are making a big mistake trying to circumvent our jurisdiction," says Peter Douglas, executive director of the California Coastal Commission, which monitors coastal development. "We will go to court."

Other states that are nervous about the environmental impact of offshore farms, such as Alaska, have sought a moratorium on development plans. In early August, MAFAC heard a spirited debate about the offshore aquaculture legislative proposal at a meeting in Juneau, Alaska. Environmental groups and state officials called for more study, and the advisory panel agreed with them. But they won't necessarily get what they want: top NMFS officials in Washington DC would only say that they are taking the group's suggestion into consideration.

High stakes

A look to the south provides a picture of the environmental issues at stake. Along the 1,600 km of Mexico's Baja California Peninsula, there are already more than a dozen offshore fish farms. The Mexican

government continues to grant permits for more, even though some businesses have shut down after setting up their nets, leaving empty pens that are a hazard for migrating sea-turtles and whales.

To stock the Mexican ranch, boats travel some 600 km down the coast to catch migrating bluefin tuna. The 35-kg fish are herded into a circular net, then slowly towed north to be anchored in deep water near Ensenada. The tow can take up to a month, during which time about 10% of the wild fish die or are lost from the nets.

Once the nets are anchored in the Pacific, farmers bring in food to fatten the tuna by about 25% before selling them, typically to Japan. Like most farmed fish, tuna are carnivorous. They are also quite picky, preferring sardines, and they are warm-blooded, which means that they require more food than cooler-blooded fish, such as salmon.

Tuna ranchers insist that their sardine-harvesting operations do not remove enough fish from the region to adversely affect the food chain. But some scientists are less sure, particularly given Mexico's patchy record for controlling its aggressive fishermen.

Near the tip of the peninsula lies Magdalena Bay, a warm-water basin that is thought to be the spawning area for much of the sardine population of the west coast of North America. After years of good management in the United States and positive climate conditions, that sardine population is at its highest in decades. But this might not last. Mexican sardine fishermen are already taking 40,000 to 60,000 tonnes of sardines a year out of Magdalena Bay, according to Mexican fishing reports. If tuna farms increase demand for sardines, the population may not withstand the pressure. "If the industry grows unchecked, it may pose a threat," says Ayayacatl Rocha Olivares, an ecologist at

the oceanographic research agency CICESE, Mexico's centre for higher education and scientific investigation in Ensenada.

Death in the water

Another problem may arise when local wildlife comes into conflict with the commercial interests of a ranch. It seems a few wayward ranch-workers have been known to take extreme measures to protect their tuna. With each hefty fish worth \$400 to \$700, a single bite from a hungry sea lion can spell a huge loss. So some took to shooting the beasts — known as *los lobos*, or wolves — that dared nip at their charges.

Residents of the coastal hamlet of Salsipuedes, for example, have complained that riflemen in skiffs regularly shot sea lions at nearby pens. Haksong Lee, the manager of pens operated by Aquaculture of Baja California, acknowledges that some shooting has happened in the past but says the practice was halted after higher nets were installed to thwart the pesky mammals. The Mexican environmental protection agency launched an investigation of the practice after enquiries by *Nature*, but so far it has not made a case against anyone.

A third concern about the tuna ranches has become apparent, thanks to farms in Australia: disease.

Back in 1995, a herpes virus hit southern Australian waters close to some tuna farms. The virus whipped across the ocean like a brushfire front, moving at 30 km a day and leaving behind it a sea of dead fish. Eventually, it was estimated that 75% of pilchards in the region died. Seabirds, from Australasian gannets to penguins, starved in the wake of the disaster^{1,2}. In 1998, another virus attack knocked out many of the remaining pilchards.

Although no one has been able to prove which factors unleashed this virus, some say it came from frozen sardines or pilchards

"The ranchers are making a big mistake in trying to circumvent California's jurisdiction. We will go to court."

— Peter Douglas

imported for the farms. "In 1995, the source of pilchards was wherever the deal was cheap," says Brian Jones, senior pathologist with Western Australia's Department of Fisheries. "They were coming from all over the world." And it was a largely unregulated trade, he says.

Tim Ward and his colleagues at the South Australian Research and Development Institute in Adelaide say the practice may have "facilitated the range shifts of pathogens that have been associated with the increased frequency of mass mortalities due to disease"³.

But no one has been able to prove where the offending 1995 virus came from. Jones and his colleagues have isolated and characterized the herpes virus from the dead pilchards. So far, the sequence matches no known herpes virus in fish elsewhere in the world.

Brian Jeffries, director of the Tuna Boat Owners Association of South Australia, a trade group for a dozen tuna-ranching groups, denies that imported fish had anything to do with the pilchard die-off. Since the 1995 outbreak, he says, bait-fish are regularly tested, and none has been found to carry a virus. Tuna ranches prefer to blame other sources, such as leaks from the water used as ballast in passing ships.

If the bait was the origin of the virus, then there is cause to worry. More than 55,000 tonnes of bait-fish were fed to Australian farmed tuna last year; about 20% of this was imported from California, says Jeffries, and another 10% imported from elsewhere. Scientists on both sides of the Pacific are watching the California sardine imports closely, because of a disease in those waters.

Mexican virus

A haemorrhagic septicaemia fish virus seems to be spreading up the west coast of North America, as the expanding sardine population migrates north from Mexican waters in search of food. Ronald Hedrick, who studies fish health at the University of California, Davis, has tracked this virus and says it appears to thrive in colder waters, where fish may be stressed from the low temperatures. There was a 58% prevalence of infection during a massive sardine die-off in the cold waters of Vancouver over 1998–99, for example⁴. "Under the right conditions, this haemorrhagic virus can contribute significantly to marine mortalities," says Hedrick.

Disturbingly, Hedrick notes, the virus is being found in more and more species: in Alaska, the virus has been linked to a die-off of Pacific herring⁵. It is unclear how fast or how far such a virus is capable of spreading, he says, emphasizing the need to keep track of international shipments of frozen bait-fish.

Jones thinks Australia has managed to dodge the haemorrhagic virus so far because the waters there are too warm. But he remains worried. "A mass die-off can happen



Nip and tuck: ranchers have been accused of shooting sea lions for taking bites out of valuable tuna.



Vast quantities of sardines (top) are frozen and shipped to farms to satisfy the picky bluefin.

again," he says. And if viruses can travel one way, they can travel the other — opening more offshore farms in the United States will only open more opportunities for unregulated trade to spread disease.

Not everyone suspects that offshore aquaculture will end with mass disease and environmental problems. Some point out that, with the nets anchored out in open water, pollution will neither accumulate nor harm sensitive coastal systems. Others go so far as to say that even coastal fish farms do not do as much harm as some people claim. Although many scientists see fish pens as a source of disease for wild species, marine ecologist Donald Kent is not so sure. "Maybe the wild fish gave lice to the penned fish," says Kent. "That is just as likely a scenario."

Kent's view is held by a minority in the scientific community. But his opinions are

being heard at high levels: he is chairman of MAFAC's aquaculture subcommittee, which advises the NMFS on policies such as the proposed offshore farming legislation. Kent is also president of the Hubbs institute, which is playing such a major part in the proposed farm on the disused oil platform. Kent sees his institute's programme as a way to meet national economic goals for farm-raised fish. "The detractors are missing the point," says Kent. "The demand for seafood is increasing; it's not going to go away."

Any experience with offshore farms in the United States is limited. There have been a few marine demonstration projects, but most were close to shore and raised fish on a small scale, making them poor tests for offshore ranches. The one pilot programme that did take place in waters far out in the Exclusive Economic Zone, in the Gulf of Mexico off Alabama, was wiped out by a storm.

This leaves proponents unable to convince critics that offshore farms would be environmentally benign. They can only point to the economic successes of other tuna farms, and emphasize that an expansion into deep waters should help the United States compete in the global fish market.

The lack of experience likewise leaves critics unable to convince the authorities that deep-water fish farms will be a disaster. But environmental watchdogs, such as the Institute for Agriculture and Trade Policy, based in Minnesota, continue to argue that these dreams of economic success court environmental trouble. If others make money out of such farms, they point out, perhaps this is because of lax rules that allow for short cuts in their management — which is precisely the sort of situation that creates environmental fall-out.

Rex Dalton is Nature's US West Coast correspondent.

1. Bunce, A. & Norman, F. I. *Mar. Freshwat. Res.* **51**, 349–353 (2000).
2. Dann, P. et al. *Mar. Freshwat. Res.* **51**, 355–362 (2000).
3. Ward, T. M. et al. *J. Mar. Sci.* **58**, 865–875 (2001).
4. Hedrick, R. P. et al. *Dis. Aquat. Org.* **55**, 211–220 (2003).
5. Marty, G. D. et al. *Dis. Aquat. Org.* **32**, 14–40 (1998).

Consumer group replies to attack on organic food

In a true scientific debate, both sides are allowed to put their cases and answer criticism.

Sir—We were disappointed that your News story “US chemist attacks consumer magazine’s food safety work” (*Nature* **431**, 117; 2004), reporting Joseph Rosen’s allegations against Consumers Union, did not supply readers with the full context surrounding Rosen’s presentation. This was part of a one-day American Chemical Society (ACS) symposium organized by Rosen on the question “Is organic food healthier than conventional food?”

Not only did Rosen’s presentation contain misleading assertions and assumptions, but we were not even allowed to ask questions after the session. Had we been invited to participate in the session or to respond to the public criticisms levelled at our work, the ACS could have engaged in a true scientific debate.

As it was, the ACS symposium lacked any alternative perspectives from consumer, environmental or organic organizations,

including the National Organic Program (a government programme that adds credibility to environmental farming and practices that have been around for decades).

As a result, no one in that session heard the broader context surrounding organic food. No one heard that the risks posed by pesticides on fruits and vegetables were not invented by Consumers Union, as Rosen led the audience to believe. In fact, by pursuing this research we were fulfilling a responsibility to our readers and the American public to evaluate the effectiveness of the Food Quality Protection Act—a 1996 law that Rosen failed to mention, aimed at protecting children.

And no one heard that—far from being a disinterested, unbiased academic, who just recently felt compelled to speak out—Rosen has for more than 15 years polarized complex scientific debates by

attacking Consumers Union (an independent, non-profit organization that accepts no advertising and has no stake in the outcome of its findings) for daring even to question the safety of the food supply. Contrary to what he told *Nature*, Rosen started criticizing Consumers Union long before we published our report on irradiation in 2003.

He has done so not only as an academic but as an adviser to the American Council on Science and Health. This is not simply “a lobby group generally supportive of the food industry”, as your News story says. It is worth noting that this organization, while claiming to represent the public interest, receives significant funding from companies whose profit margins depend on the continued use of pesticides.

Urvashi Rangan, Jennifer Schecter

Consumers Union, 101 Truman Avenue, Yonkers, New York 10703, USA

Meyer case poses a challenge to the system

Sir—Your News story “Junior biologists score partial victory over lab conditions” (*Nature* **430**, 7; 2004) illustrates well the extreme difficulties of dealing with high-profile scientific misconduct cases.

However, as some of the junior scientists involved in this case, we feel that your story may have left the unfortunate impression of a conflict largely over lab conditions and management style.

In fact, Axel Meyer was declared guilty of scientific misconduct on eight out of a sample of 13 counts documented by an independent university commission, who followed criteria set by the main German scientific bodies.

It is time we recognized that scientific misconduct is not only about data manipulation. Of the eight confirmed allegations, the four that concern grant plagiarism and authorship manipulation should be taken particularly seriously by the scientific community. The original complaint presented to the commission does more than “hint” at scientific misconduct—it also records damage to junior researchers’ career prospects.

The complexity of this case poses a difficult challenge to the German academic system, and it is not surprising that all institutions involved are taking their time to decide about the consequences to be imposed. For the 16 junior scientists who overcame institutional resistance to defend

their rights, it is important that verdicts are respected and suitable measures are applied to ensure and protect scientific integrity. Until then, a partial victory remains just a moral victory.

Edward Málaga-Trillo*, Gabriele Gerlach†

**Department of Biology, University of Konstanz, 78457 Konstanz, Germany*

†Marine Biological Laboratory, 7MBL Street, Woods Hole, Massachusetts 02543, USA

Other signatories of this letter:

Julia Day, Zoltán Ivics, Zsuzsanna Izsvák, Eric Rivera-Milla, Simona Santini, Jens Seckinger

Meyer: disagreements but no misconduct

Sir—Your News story “Junior biologists score partial victory over lab conditions” (*Nature* **430**, 7; 2004) reported accusations by 16 former postdocs and graduate students against Professor Axel Meyer at the University of Konstanz, Germany. The complaints mainly concerned laboratory conditions, not the quality of research. As reported in *Nature*, an investigation by a university committee rejected some of the allegations but accepted others.

As former students, postdocs and scientific collaborators, we were surprised by the committee’s decision to accept some of the complaints and by the ensuing media coverage. Our own experiences of working in or collaborating with Axel Meyer’s lab have been far more positive. Those of us who have published with

Meyer found that he provided crucial intellectual contributions to manuscripts. On numerous papers from his lab he was not an author.

Working in Meyer’s lab, like in many others, demands a high level of dedication and it is sometimes necessary to change the direction of research to ensure these standards. This is a common situation in labs where scientific progress is the foremost priority.

Meyer contributed productively to the selection and planning of research projects while allowing associates sufficient independence when the research was proceeding well. Technical, logistical and financial support were always available to all lab members. Collaboration with Meyer took place in a fair, open and cooperative atmosphere.

Although some of us, at times, have disagreed strongly with him, we have all benefited scientifically from our past experiences or interactions with Meyer’s group and unanimously support him now.

Miguel Vences*, Rafael Zardoya†

**University of Amsterdam, PO Box 94766, 1090 AT Amsterdam, The Netherlands*

†Museo Nacional de Ciencias Naturales, J. Gutiérrez Abascal 2, 28006 Madrid, Spain

Other signatories of this letter:

E. Abouheif, M.-A. Akimenko, F. Andreone, M. L. Arnold, L. Bargelloni, J. L. Boore, H. Brinkmann, R. Cerff, M. P. Cummings, S. Edwards, M. Ekker, R. Grosberg, I. Horovitz, D. M. Lambert, M. Lang, P. Lockhart, T. Mack, M. Mari-Beffa, D. P. Mindell, R. Owino, T. Patarnello, D. Penny, D. Reznick, P. A. Ritchie, L. Rüber, M. Scharf, C. Sturmhuber, J. S. Taylor, Y. Van de Peer, E. Verheyen, M. Veith, D. R. Vieites, J.-N. Volff, J. Wittbrodt

Alert to a European epidemic

Funds must be forthcoming for an effective EU Centre for Disease Control.

S. Ragnar Norrby

The eruption of severe acute respiratory syndrome (SARS) in 2003 was characterized by the rapid spread of the outbreak across several countries around the world. Although the World Health Organization (WHO) provided extremely fast information and support to afflicted countries, the European network of state epidemiologists and national agencies failed to respond quickly enough. Important information about possible cases of SARS or European travel restrictions were often communicated with delays of 48 hours or more. European healthcare officials hope to avoid repeats of such inadequate responses, and are determined that faster coordination and better surveillance will be made possible by a European Centre for Disease Control (ECDC), similar to the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia. A decision about the new ECDC director is expected shortly, but it is already clear that this director will face an uphill struggle to create a European equivalent to the CDC.

Uncontrollable outbreaks of infectious diseases are a public-health crisis waiting to happen. New and emerging infections, including SARS and HIV/AIDS have appeared, and older infections, such as tuberculosis and malaria, are far from being eliminated. Resistance to antibiotics continues to increase, with no new drugs on the horizon. We also face threats from large-scale outbreaks of infectious diseases, especially an influenza pandemic, and from the deliberate release of infectious agents by criminals or terrorist groups. Few of these problems are being addressed in a constructive way — industry is failing to develop new antibiotics and vaccine production is insufficient to meet the needs of a pandemic.

Against this background, it seems clear that Europe would benefit from a regional organization similar to the Atlanta CDC, which has large resources in terms of scientific expertise and laboratory capacity, and which can deploy field forces (epidemiologists and laboratory equipment) to the site of an outbreak at very short notice. The current difficulties of the European Union in collating and disseminating epidemiological information — as seen with SARS — will only get worse with the addition of ten new member states. Therefore the creation of an independent agency for surveillance and control of communicable diseases has been welcomed, and from 2005 the ECDC will operate from a base in Stockholm, Sweden.



Staff trained in lab work are in short supply.

But there remain fears that the ECDC has been structured in a way that leaves it powerless to effect change. It will lack both regulatory authority and laboratory resources — which will curtail its independence. Lab resources are considered by many to be a necessity for modern infectious diseases epidemiology. But this particular omission could be corrected if the third weakness, a small budget, is improved over time.

Why does Europe need its own CDC? On a global level, the WHO regularly communicates information from official government sources. But such information is not always complete, as governments try to balance openness with the negative effects that epidemics have on travel and tourism. In addition to the CDC — which tracks both the US and global situations — there is the

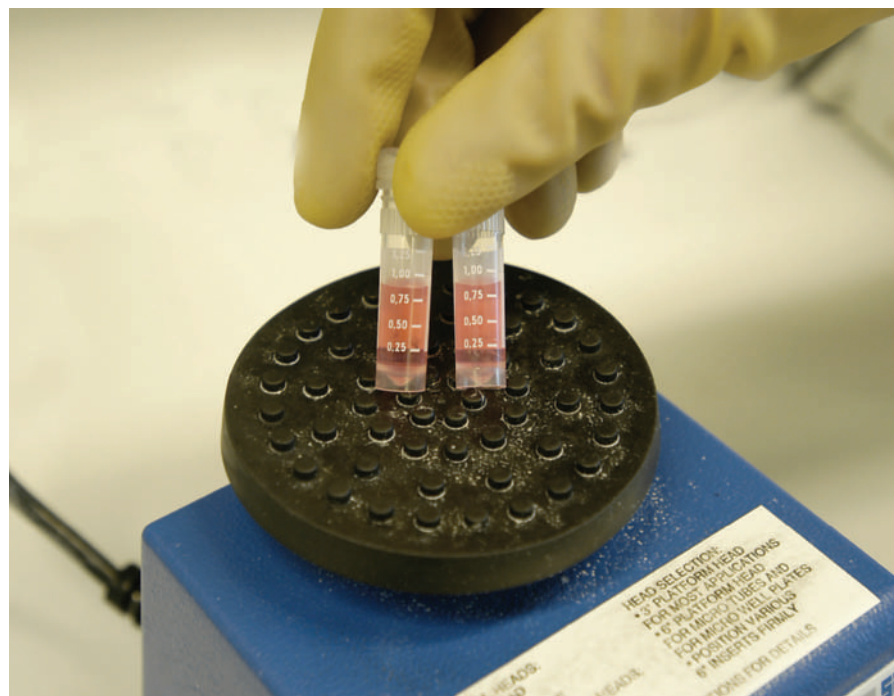
internet-based ProMed, now ten years old. This e-mail system (open to anyone who is interested) is run by the International Society for Infectious Diseases and is a major source of rapid and reliable information. One advantage of the Atlanta CDC is its ability to combine surveillance with active field forces. At present, Europeans have to call the CDC or the WHO for help when there are outbreaks in Europe. Although SARS was stopped in its tracks by such global coordination, Europe may not be so lucky next time.

One of the main tasks of the ECDC is to “search for, collect, collate, evaluate and disseminate relevant scientific and technical data”¹. This is currently carried out on a national level in Europe. But EU countries vary considerably in terms of resources and which diseases are notifiable. The surveillance systems and the quality of the data generated also vary. For example, each year more cases of salmonellosis are reported for Swedish tourists visiting some European countries than for the entire native populations of those countries. Standardization of these systems is not likely to occur in the near future, except for the most severe infections, which will be regulated by the new International Health Regulations, to be ratified by the WHO in 2005.

Modern epidemiological investigations go beyond collection and analysis of clinical reports, often requiring sophisticated laboratory studies of the microorganisms — molecular epidemiology — to achieve reliable analyses. Accordingly, CDC and several European government agencies are given access to large laboratory facilities. But the ECDC’s preliminary budget for 2005–2007 does not include funds for lab activities within the centre itself, nor will there be sufficient funds to pay for more than limited services at national laboratories.

Widespread outbreaks of communicable diseases also require access to epidemiological expertise that can be rapidly deployed in the field. Containment of the 2003 SARS outbreak was principally due to the WHO’s rapid deployment of advisers to manage the epidemic on site. The WHO and CDC have also helped control many other outbreaks, including Ebola in Africa. Europe lacks coordinated resources of this kind.

Together, the relatively small ECDC budget and lack of facilities will prohibit the creation and maintenance of a European field force. They will also limit the independence of the centre and will make it harder to recruit the most competent staff. In the short term, locating the ECDC close to existing facilities in Stockholm may improve this



situation. In addition to the excellent facilities at the Karolinska Institute, the Swedish Institute for Infectious Disease Control has an advanced biosecurity (biosafety level 4, BSL4) laboratory. The new ECDC director should be able to allocate some funds for lab work, subject to approval by the management board.

What ECDC can, and hopefully will, do is train members of an international task force and coordinate their activities within or outside Europe. Because of its small allocated budget, salaries for such staff members can be covered by ECDC only when an outbreak occurs. Between outbreaks, members of the force must be supported by their home countries.

The ECDC will have no regulatory power¹ in accordance with the rule that the European Union does not issue directives in the field of public health. This is why the choice of director is so important. The director must be a leader who can attract epidemiologists, microbiologists and scientists of very high standards. The tasks of the ECDC are formulated in such a general way that the

The problems surrounding resistance are heightened by a lack of new antibiotics — we are close to the point where we might return to the pre-antibiotic era³. In addition to greater morbidity and mortality, resistance increases costs through prolonged hospital stays and the use of more costly drugs.

One of the most worrying threats is a new influenza pandemic, whose likelihood increases with every outbreak of avian flu in Southeast Asia and Europe. Of particular

“We also face threats from large-scale outbreaks of infectious diseases, especially an influenza pandemic, and from the deliberate release of infectious agents by criminals or terrorist groups.”

epidemiological field force within its allocated budget. But with the expectation of future funding increases, the ECDC should initiate field forces within the member states. A substantial boost to the ECDC would be money well spent as the costs of antibiotic resistance, let alone of SARS or influenza outbreaks, are likely to be enormous. ■

Competing interest: The author is director-general of the Swedish Institute for Infectious Disease Control, an expert government agency for surveillance, control and research in the field of communicable diseases. Citizens from the host country are excluded from the ECDC director position.

Play it again, John

A look back at the birth of game theory some 60 years ago.

Theory of Games and Economic Behavior: Sixtieth-Anniversary Edition

by John von Neumann & Oskar Morgenstern
Princeton University Press: 2004. 704 pp.
\$50, £32.50

Karl Sigmund

"Posterity may regard this book as one of the major scientific achievements of the first half of the twentieth century." So began a 1945 review of this book in the *Bulletin of the American Mathematical Society*. There followed a rash of equally enthusiastic articles, including one on the front page of the Sunday issue of *The New York Times* — enough to make any publisher walk on air. This sixtieth-anniversary edition of *Theory of Games and Economic Behavior* is a fitting sign of lasting gratitude from Princeton University Press.

As well as being an achievement in the first half of the twentieth century, the book proved a pacemaker for its second half. A 600-page monograph usually represents the summing-up of a mature scientific field. In this case it was the birth cry of game theory. The theory is no great help for playing poker, but uses terms such as 'player', 'move' and 'pay-off' to analyse social interactions.

The field developed in ways that were hardly anticipated by the book's authors, John von Neumann and Oskar Morgenstern. In 1950, a 20-year-old called John Nash introduced a fundamental equilibrium concept, daring, as he later said, "to deviate from the line (in the sense of party line)" laid down by the two founders of the field. But the economic establishment remained aloof for some time. At first, the field grew mostly through the work of brilliant young mathematicians with only a minimal background in economics, and ideas that were very much their own. One of them, Harold Kuhn, has written the introduction to this reissued edition.

One of the main parts of the book, the theory of two-person zero-sum games, analyses situations of pure competition, which later were felt to be of only marginal relevance to most economic interactions. But this theory, which was mathematically equivalent to linear programming, became one of the hottest topics of the 1950s, leading to several Nobel Prizes in economics. Another main section of the book is devoted to a theory of coalitions, centred on a solution concept that, as it later turned out, could not always be fulfilled. Von Neumann and Morgenstern had stated that "such an



Game for anything: it's hard to guess what people will do, as human behaviour is often irrational.

eventuality would certainly necessitate a fundamental change in the theory".

Moreover, von Neumann and Morgenstern had intended a "General theory of rational behavior" (the original book title, fortunately dropped), prescribing what rational agents ought to do. It was based on an axiomatic treatment of personal utility that was plausible but was later shown to be in patent contradiction to actual human behaviour. As Ariel Rubinstein writes in an afterword to this anniversary edition, game theory does not have normative implications. Rather, it is a cousin of logic, a mathematical tool more useful for description than for prediction. Today, game theory is used routinely in evolutionary biology, helping to analyse societies of animals or bacteria. In the same vein, experimental economics based on simple games has grown tremendously, but without finding much rationality.

The value of the book is largely historical today, having been superseded in most classrooms by Duncan Luce and Howard Raiffa's *Games and Decisions*, which appeared in 1957. But this history is exciting. What did von Neumann and Morgenstern have in common, apart from an Austro-Hungarian background and a neighbourhood in Princeton? 'Good-time Johnny', as von

Neumann was known, a Wunderkind come to joyful maturity, was universally hailed as a mathematical genius of amazing versatility. He wrote ground-breaking papers on formal logic, quantum theory, functional analysis and group representations. Morgenstern, a maverick economist, struggled all his life with elementary mathematics but never gave up.

It is said that von Neumann was both the father and the mother of the book, and Morgenstern its midwife. During the 1940s, von Neumann launched into a stupendous variety of mathematical contributions to ballistics, sea mines, thermonuclear explosions and programmable computers. The general view is that, in between his restless travels, he occasionally breezed into Princeton with a handful of notes on game theory, explained them (in German) to Morgenstern, far into the night, and left him with the task of typing it all up. But Morgenstern's article describing the collaboration, and particularly his diaries — kept at Duke University in Durham, North Carolina, and analysed by historian of economics Robert Leonard — suggest that the meeting of minds went deeper than that.

In 1928, von Neumann had published in Berlin a paper on parlour games. Meanwhile Morgenstern, in Vienna, published a book entitled *Wirtschaftsprognose*, in which he

claimed that economic predictions are in principle inconsistent, because they cause agents to act in a different way from that predicted. His favourite example was the fictional detective Sherlock Holmes being pursued by the murderous Moriarty and having to decide where to leave a train, with Moriarty anticipating this, in a vicious circle of mutual outguessing. Morgenstern, the director of a forecasting agency in Austria, saw in this infinite regress ("He thinks that I think that he thinks...") a basic obstacle to forecasting the decisions of interacting individuals, no less fundamental than the uncertainty principle or the incompleteness theorem. Only later did he realize that von Neumann's 'maximin' solution, yielding the highest guaranteed payoff, offered a way out: mixed strategies (such as 'leave at Dover with probability of two-thirds') can be optimal, even if an adversary manages to guess them.

Although Morgenstern's love affair with mathematics was purely one-sided, he was closely acquainted with some of Vienna's best mathematicians, including the geometer Karl Menger, who co-discovered dimension theory, and the logician Kurt Gödel. Menger was so distressed by the political turmoil of the time that in autumn 1934 he withdrew to a mountain resort, where he wrote an odd little booklet on mathematical ethics, based on the 'tolerance principle'. Just as there is not one geometry but many (euclidean, hyperbolic and elliptic, for example), so there is not one system of moral norms but many, from which individuals make their own choice. Menger then studied how groups of adherents to different norms interrelated.

In 1938, Morgenstern was prevented from returning to Austria from the United States by the Nazis, who had blacklisted him. During his exile in Princeton, he sought to extend Menger's ideas in a paper on "Maxims of behavior", and to analyse societies of agents whose decisions impinge on each other. This was when von Neumann, who had not worked on game theory for ten years, stepped in, and eventually suggested that they write the paper together, for the benefit of economists. "Here was my gift from heaven," wrote Morgenstern. Their manuscript started growing relentlessly, becoming first a two-part paper, then a small pamphlet that Princeton University Press agreed to publish, and then the authors "completely forgot about any restriction to 100 pages".

The final product is not for the faint-hearted, and few will have had the stamina to work through its pages crammed with footnotes and formulae. But by its mere existence, the heavy tome marked a turning point in economics, challenging it to become a mathematical discipline at last. ■

Karl Sigmund is in the Faculty of Mathematics, University of Vienna, Nordbergstrasse 15, 1090 Vienna, Austria, and the Institute for Applied Systems Analysis, Laxenburg, Austria.



JEFFREY L. ROTMAN/CORBIS

Into the depths: the body can muster an extraordinary range of defences against extreme conditions.

Life on the edge

The Biology of Human Survival: Life and Death in Extreme Environments

by Claude A. Piantadosi
Oxford University Press: 2003. 280 pp.
£24.95, \$35

Mike Stroud

Stories of human achievements and survival against the odds have always been fascinating. Whether in the context of simply living in the world's harshest environments, mounting expeditions to its hottest, coldest, highest or deepest places, or coping with the aftermath of disaster, everybody wonders at just how the body copes. Answers lie in the study of environmental physiology, the responses and adaptations that can take men and women to extremes.

The Biology of Human Survival is an extraordinary environmental physiology text. The topics covered range far beyond biology to include the physics and function of artificial aids that allow humans to cope with extremely hostile environments. But engineering approaches are not just used to describe life-supporting technologies — the author also uses them to explain biological concepts. This approach helped me to understand some concepts that I had previously struggled with. Occasionally, however, the opposite applies.

The book begins by describing the limits to the range of environments that can support human life, along with the principles of survival, adaptation and life-support

systems. The historical background to environmental physiology is fascinating, but as the book moves on to adaptation (physiological changes in response to environmental stress) and maladaptation (adverse changes resulting from adaptation), some topics were unclear and others were made unnecessarily complex. For example, the author stresses the importance of discriminating between technical definitions, such as adaptation, acclimatization, acclimation, accommodation and habituation, but then, I feel, blurs the boundaries. He has also focused particularly on maladaptation and cross-acclimation (adaptive changes to one type of environmental stress that prove beneficial during exposure to stress from a different type of environment), perhaps ascribing more importance to these processes than they deserve. Indeed, he suggests that adverse effects of cross-acclimation between cold and hypoxic responses contribute to the difficulty of climbing Mount Everest in winter. In reality, this must be insignificant compared to winter's cold, storms and jet-stream winds.

But back to the book's strengths. Several chapters cover adaptation to heat and cold in detail. There are lengthy descriptions of human responses and adaptation to icy environments, but our physiological responses to heat (which are far more effective than those to cold) are not covered in such depth. This book is not, then, a definitive work on environmental physiology. But viewed as a collection of thought-provoking pieces about this field it becomes a tour de force. This is especially true when the author strays from his title, covering not just engineering and biology, but also life that is far from

Science in culture

Great, not gruesome

Pat York's photographs of dissected humans represent a fine body of work.

Martin Kemp

The notable eighteenth-century German anatomist Bernhard Siegfried Albinus demonstrated that skin colour — the characteristic we use most readily to judge someone's ethnic origins — was literally a superficial matter. He also disclosed, not least through the stylish illustrations he commissioned from Jan Wandelaar for his grand *Tabulae sceleti et musculorum corporis humani* in 1747, that the inner topography of the dissected body was no less wondrous and beautiful than its exterior. How could it be otherwise, as anatomists had long claimed that the human frame was God's greatest achievement as divine engineer?

These themes are reincarnated in Pat York's remarkable photographs of dissections performed by the Los Angeles anatomist Marc Pick. Over the course of seven years, York has been granted regular access to the products of Pick's manual, intellectual and creative skills. She describes what she has witnessed as "one of the most awesome experiences of my life" and adds: "I feel an unwelcome affinity for these still, complex bodies."

On the face of it, such ostensibly gruesome subjects could hardly represent a sharper contrast to the images that first gained her fame, namely her portraits of celebrities, mostly from the Hollywood world of film. Yet when we look back on how she has presented the famous, and also at her suites of photographs of workers in the nude and of vibrant people over 75 years old, we can see that she has always been interested in what lies below the surface. The body, the face and the eyes have

always acted for her as a "window to the soul".

Now, in a work such as *Universal Self-Portrait*, shown here, she is re-examining our relationship with our own bodies. Superbly executed, direct and starkly compelling, the image challenges us to look again at what we all have inside ourselves but



prefer not to confront visually or emotionally. We find it more comfortable to live our lives on the surface of who we are defined as being. In her image, the brain, folded like some complex product of vast geological torsions, is the seat of much of what makes us individual, yet she emphasizes that "the body, when the skin is peeled away, has no distinctions of colour, race or religion".

The strange wonder of our inner selves draws York into the biggest issues that have confronted humanity over the ages. "We all have souls, we all have hopes and dreams, both fulfilled and

otherwise, and we all have loves and passions. Where are those souls now? Have they been reincarnated? Are they in heaven? Is there just a void?" She continues: "We all share this miraculous, complex interior — far more complicated than any technological advances we have made in our

society. I am constantly mesmerized at the complexity of the human experience, and baffled by the existence of hatred and violence. If only in life we could see our similarities — differences do not apply. It seems ironic that the only thing we are sure about in life is that we shall die, the one fact we all seem to wish to escape. *Universal Self-Portrait* is all of us."

Albinus subscribed to the traditional notion that the justification for dissection was to "know thyself". Pat York's work stands centrally in this tradition in a way that the flashy, opportunistic and exploitative displays of Gunther von Hagens do not. Von Hagens poses beautifully dissected bodies in irrelevantly rhetorical poses. The great anatomists knew that presentation, pose, significance and communication should be totally

integrated if they were to do their momentous job with the highest levels of integrity. York's pictures of the dissected body similarly allow no compromise in allying form and meaning.

York's photographs went on display this week at the Galerie Gmurzynska in Cologne, Germany, where they can be seen until 23 October.

Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK, and co-director of Wallace Kemp Artakt. His new book, *Leonardo*, was published last week by Oxford University Press.

human. The piece on the physiology of the camel in the section on salt and water is masterly, and there are fascinating descriptions of the interaction between primitive life and Earth's early atmosphere. There is also a beautifully worked analysis of why you should never drink sea water.

The chapter on nutrition and survival, although generally excellent, does perpetuate some rather outdated views. For example, it states that the main difference between the forms of malnutrition known as marasmus and kwashiorkor is in the level of protein intake; elsewhere in the book, the author mentions the more current idea that kwashiorkor and its accompanying oedema are more a product of free-radical membrane damage than low protein ingestion. There is also misleading information about the

subsequent reintroduction of normal nutrition (refeeding), and there are some rather simplistic views on vitamin deficiency. These include the idea that the main problem with vitamin A depletion is ocular, whereas we now know that vitamin A deficiency also impairs responses to infections of the gut and respiratory tract, leading to deaths in people who have very little or no eye damage.

Biology and engineering are mixed even more freely in the second half of the book than in the first. Descriptions of the technical engineering solutions to the high pressures of the deep sea and low pressures of the high mountains are balanced excellently with descriptions of pressure physiology and the illnesses that can stem from pressure change. Just as in the earlier part of the book, in which problems caused by the cold are illustrated

by compelling tales including those of Scott of the Antarctic and the *Titanic*, the author uses famous disasters to bring the issues in this section to life. The sinking of the Russian submarine *Kursk* is used to great effect.

Towards the end of the book there is a surprising but topical diversion into survival in the face of nuclear, biological and chemical weapons of mass destruction. This is a rather depressing digression, but it is both interesting and relevant.

The book ends with a section lifting us away from Earth's limitations to describe the exciting physiology and engineering of high-performance aircraft and space flight. The final chapter even speculates on the requirements for and limitations to future human colonization of other planets, and so ends on a positive note, as will I. There is no doubt

that this book will be enjoyed widely and will be much appreciated by both specialists and scientifically thoughtful lay readers. ■

Mike Stroud is at the Institute of Human Nutrition, Tremona Road, Southampton SO16 6YD, UK. His expeditions to the Earth's extremes include ultra-distance runs in the Sahara and the first unsupported walk across Antarctica.

Uncovering chromosomes

The Man Who Invented the Chromosome: A Life of Cyril Darlington

by Oren Solomon Harman
Harvard University Press: 2004. 342 pp.
\$49.95, £32.95, €46.10

Rena Selya

Cyril Dean Darlington was responsible for our understanding of the behaviour of chromosomes during mitosis and meiosis. Working with plants at the John Innes Horticultural Institution in London in the 1920s, he displayed a keen eye for microscopy, and spent much of his time exploring the structures of the cell. His scientific insights, however, were more often the result of theoretical reasoning than careful empirical observation. His 1932 masterpiece, *Recent Advances in Cytology*, earned him both great praise and harsh criticism because of his unorthodox methods.

As Oren Solomon Harman shows in *The Man Who Invented the Chromosome*, Darlington's controversial cytological research clarified many basic biological issues and provided essential evidence for the evolutionary synthesis of the 1940s. He 'invented' the chromosome by describing its behaviour in a way that made genetic and evolutionary sense. His description of the way chromosomes line up with their homologous copies before cell division settled a long-running debate among cytologists over whether chromosomes pair up end to end or next to each other, and accounted for the phenomenon of crossing over. His contributions to biology were significant, yet he has been overlooked in the history of the life sciences.

One of the book's strengths is Harman's deft description of the confusion that reigned in the biological community in the first half of the twentieth century. He shows how well-educated, talented researchers could draw opposing biological conclusions from experimental data because of their conflicting disciplinary affiliations and generational perspectives. When Darlington began his work at the John Innes, under William Bateson, he stumbled into a community in epistemological upheaval. Despite his position as one of the founders of modern genetics, Bateson



Look at it my way: the results of Darlington's unorthodox methods eventually convinced his critics.

resisted the chromosome theory of inheritance because he felt there were too many experimental exceptions for it to explain mendelian inheritance. Young US geneticists were willing to extrapolate from data from a model organism, whereas cytologists spent years accumulating evidence from a range of plant and animal species before drawing general biological conclusions. Harman gives the reader a sense of Darlington's growing confidence as he made bold claims that were eventually accepted by biologists of all disciplines.

A strong commitment to an evolutionary perspective led Darlington to some unpopular conclusions, which he published in books and articles aimed at a wide audience. Convinced that biological principles, especially genetics, dictate human values, he espoused strong eugenic programmes and argued for the biological existence of race, especially after UNESCO published its statements on race in the early 1950s. Darlington studied human history through the lens of evolutionary pressure, concluding that genetic and environmental diversity should be maintained to ensure the survival of the human race. Although his ideas were unpopular so soon after the Second World War, he felt that the time had come for science to determine morality: religion and politics should be replaced by evolutionary logic for individuals, countries and humanity.

The influence of science on society was unidirectional, however. Darlington firmly believed that political considerations should never influence science, whether under liberal or totalitarian governments. He was one of the first scientists outside the Soviet Union to recognize the danger in Trofim Lysenko's scientific and political positions. Despite the fact that some of Darlington's work on cytoplasmic inheritance could have supported a

lamarckian view of heredity, he criticized Lysenko's science while other biologists played down its influence. He took no pleasure in accurately predicting the terrible fate of Soviet geneticists, and he chastised colleagues who were loyal to the Communist party.

Darlington was a lifelong diarist, and Harman makes fine use of the red bound notebooks that now reside in the Bodleian Library in Oxford. He attributes Darlington's scientific success, after a lonely and academically undistinguished childhood, to a combination of intelligence, arrogance and the desire to please a demanding, emotionally distant father. Harman chronicles Darlington's tumultuous personal life (he had three wives), and incorporates the recollections of two of his children. He describes Darlington's thoughts and feelings in a novelistic manner, so the scientist comes across as a complex, if not altogether likeable, person.

The prose is occasionally melodramatic: "It was as if the chromosomes themselves, at the other end of the ocular lens, could feel it: Darlington was hungry." Still, the style conveys Darlington's human side well. Harman does not fall into the trap of tediously chronicling the life of his subject, but rather presents Darlington's scientific research and popular writings as the expression of paradoxical personal and intellectual themes.

Because of his controversial views and brusque personality, Darlington faded from the public eye before his death in 1981. But the spectres of genetic determinism and political interference in science remain with us, and Harman provides a cautionary tale for those who seek to tie our humanity too closely to what is found in our chromosomes. ■

Rena Selya is in the Department of the History of Science, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, USA.

The domino effect

When taking a risk proved a wise choice for one postdoc.

Jamshed Tata

Job security is a rare luxury for postdocs. But early in my scientific career, I turned down a secure position to follow a challenge from a senior scientist. My gamble paid off, and a quick succession of unexpected results enabled me to return to a secure career path.

When doing my second postdoc in 1959 at the National Institute for Medical Research (NIMR) in London, I attended a joint meeting of the British and Scandinavian biochemical societies in Turku, Finland. There, I listened to Lars (Lasse) Ernster from the Wenner-Gren Institute of Stockholm University talk about the uncoupling of oxidative phosphorylation by L-thyroxine — the precursor to active thyroid hormone — when added in large amounts to isolated mitochondria. Thyroid hormone was known to increase the basal metabolic rate (BMR), the rate at which a warm-blooded animal consumes oxygen and other fuels at rest. At that time, Ernster and many others believed that thyroid hormone's physiological action could be explained by these *in vitro* experiments, which showed that the hormone uncoupled the consumption of oxygen from energy (ATP) production. After I told him — with some chutzpah — that I was not convinced that is how the hormone would work *in vivo*, he responded, “if that is so, come and prove it in my laboratory”.

This challenge was enticing, but put me in a dilemma. I had been offered a rare Senior Beit Fellowship to continue at NIMR. Turning down this offer and accepting Ernster's would be risky. What if I turned out to be wrong?

But taking a gamble, I took up Ernster's offer and arrived in Sweden a year later. Luckily, the very first experiment indicated that I was on the right track. We soon demonstrated that under physiological conditions, triiodo-L-thyronine (T_3) — the active form of thyroid hormone — regulated the amount of ATP made by mitochondria in liver and skeletal muscle in strict proportion to their rate of respiration (that is, respiration and energy production remained coupled). The result was in stark contrast to the uncoupling previously obtained by the Wenner-Gren and other groups with highly toxic doses of the hormone *in vivo* or *in vitro*. The next set of experiments most gratifyingly explained why. T_3 caused a net, tissue-specific increase in the content of mitochondrial enzymes, particularly dehydrogenases (the first step of fuel burning for energy production) and membrane proteins



The author (left) meets Fritz Lipmann.

involved in oxidative phosphorylation. Then I had more luck — I was able to demonstrate the marked effect of T_3 on protein synthesis by ribosomes and microsomes that had been isolated simultaneously with mitochondria from the same tissues. More importantly, this was manifest with a reduced latent period for the enhancement of mitochondrial respiration, which was significant as inhibition of protein synthesis blocked the stimulation of mitochondrial respiration and BMR by T_3 . In a paper in *Nature*, followed by a more detailed account in the *Biochemical Journal*, we dismissed the uncoupling hypothesis and proposed that thyroid hormone regulated the body's BMR by a selective control of the synthesis of key mitochondrial enzymes and proteins. Ernster not only graciously accepted this conclusion and co-authored these and subsequent publications, but also wrote about my findings to several proponents of the uncoupling hypothesis, urging them to listen to my arguments against it when I visited the United States the next year.

In 1961, I toured the United States explaining our results, with varying degrees of success, to eminent groups that championed the uncoupling hypothesis, such as those of Albert Lehninger and Henry Lardy. My last port of call was The Rockefeller University, New York, where — to my great surprise and joy — Fritz Lipmann, one of the first proponents of the uncoupling of oxidative phosphorylation, encouraged me by generously sharing some unpublished lab notes from the mid-1950s that described a stimulatory effect of thyroxine on protein synthesis. Back in Stockholm, I met Peter

Medawar at a reception at the Karolinska Institute on the day after his Nobel prize award. He was soon to take up the directorship of NIMR, and invited me to return there to continue the work I had started in Sweden. The year 1961 was truly eventful, not just for my future career — our third child was born just before the stroke of midnight on New Year's Day, and I escaped a fatal air accident when on my United States tour!

I went back to NIMR as a junior member of the staff on a three-year contract at a time when biochemistry and molecular biology were rapidly moving forward. Although attempting to determine if the latency in the response of protein synthesis to T_3 came after stimulation of transcriptional activity, Chris Widnell (my first graduate student) and I unexpectedly stumbled on the multiplicity of eukaryotic RNA polymerases, one synthesizing ribosomal RNA (rRNA) and the other messenger RNA (mRNA). Our paper on the novelty of multiple enzymes was at first rejected by the *Journal of Molecular Biology* with a curt note saying that “since bacteria have a single RNA polymerase there is no reason to accept that animal cells should have more”! (The paper was published in another journal.) Later, the groups of Bob Roeder (University of Washington, Seattle) and Pierre Chambon (Institut de Chimie Biologique, Faculté de Médecine, Strasbourg) resolved three polymerases synthesizing different classes of RNA by exploiting the enzymes' differential inhibition by the fungal toxin α -amanitin. But one should not be too greedy for luck, as Widnell and I soon showed that T_3 did stimulate RNA polymerase II and mRNA synthesis, preceding that of protein synthesis. The kinetics of stimulation of nuclear RNA polymerase II could be superimposed on the accumulation of T_3 by nuclei — undoubtedly reflecting the binding of the hormone to nuclear T_3 receptor. The intimate association between the steroid/retinoid/ T_3 receptor superfamily and regulation of transcription is now part of the established dogma of ‘receptorology’.

Clearly, the risky move, followed by generous help and inspiration from some leaders in my field, made this my turning point. Within a period of three years, the dominoes had fallen down correctly and in rapid succession. Had I been obliged to apply for separate grants at each point, or taken the sure and safe route, this would not have been possible. I was truly a most lucky postdoc! ■

Jamshed Tata is at the National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK.

Electrons hang ten on laser wake

Thomas Katsouleas

Electrons can be accelerated by making them surf a laser-driven plasma wave. High acceleration rates, and now the production of well-populated, high-quality beams, signal the potential of this table-top technology.

Huge particle accelerators have been at the vanguard of research in particle physics for more than half a century; through high-energy collisions of accelerated particles, the fundamental building-blocks and forces of nature have been revealed. The latest project, the Large Hadron Collider (LHC) currently under construction at CERN in Geneva, will attempt to find the Higgs boson, a particle associated with the mechanism through which all other known particles are thought to acquire their masses. But the size and cost of such machines — for the LHC, a 27-km circumference and several billion euros — are fuelling a serious effort to develop new and more compact accelerator technologies. Three reports^{1–3} in this issue (from page 535) announce fresh progress, using a principle known as plasma wakefield acceleration.

Plasmas — gaseous ‘soups’ of dissociated electrons and ions — offer a means of acceleration that could be realized on a table top⁴. Waves can be generated in a plasma using short laser pulses; electrons or their antimatter counterparts, positrons, can then ‘surf’ the electric field of a wave’s wake. Particles have been accelerated in wakefields at rates that are more than a thousand times higher than those achieved in accelerators based on conventional large-scale technology. However, whether plasma wake-

field accelerators could produce the high quality of beam needed for applications in high-energy physics, and in other areas of research and medicine, remained in question. The results now presented by Geddes *et al.*¹, Mangles *et al.*² and Faure *et al.*³ are a milestone in this regard. They provide the first demonstration that a beam of electrons can be accelerated in a wakefield to a single energy. Moreover, their beams are of high quality (having a small angular divergence) and significant charge (about 10^9 electrons).

In a conventional accelerator, charged particles such as electrons, protons or their antiparticles are accelerated by an alternating, radio-frequency electric field through long metallic cavities (around a metre long for medical applications, but several kilometres long for high-energy physics). The rate of acceleration is limited by the peak power of the radio-frequency source and, ultimately, by electrical breakdown at the metal walls of the accelerator. Laser-driven plasma waves overcome both of these limitations: the high peak power of lasers is unmatched, and the plasma, as it is already an ionized gas, is impervious to electrical breakdown. In 1995, Modena *et al.*⁵ made clear the remarkable potential of this scheme, and it has been confirmed by subsequent experiments. Using the radiation pressure of a laser

to drive a compressive oscillation in the plasma (like a sound wave, but with electrostatic repulsion rather than pressure as the restoring force), electrons have been accelerated from rest to an energy of 100 megaelectronvolts (MeV) within a distance of 1 mm — more than 5,000 times shorter than the distance required to reach that energy in a conventional accelerator.

But acceleration rate is only one measure of a good accelerator. The number of particles in a beam, and their spread in angle and energy, also matter. In 2002, Malka *et al.*⁶ showed that well-collimated beams of 10^8 electrons could be produced within an angular spread of 3° by a laser-driven wakefield; in these experiments, however, the energy spread of the beams was 100%. This wide range of energies occurred because the particles were trapped from the background plasma — in much the same way that white-water gets trapped and accelerated in an ocean wave — rather than injected into a single location near the peak of the wave (as is done in a conventional accelerator). But injection is difficult in a wakefield accelerator because the wavelength of the plasma wave is tiny — typically 10,000 times shorter than the usual 10-cm wavelengths of the radio-frequency fields in conventional accelerators. Successfully injecting tightly packed bunches of particles near the plasma-wave



TIME LIFE PICTURES/GETTY IMAGES

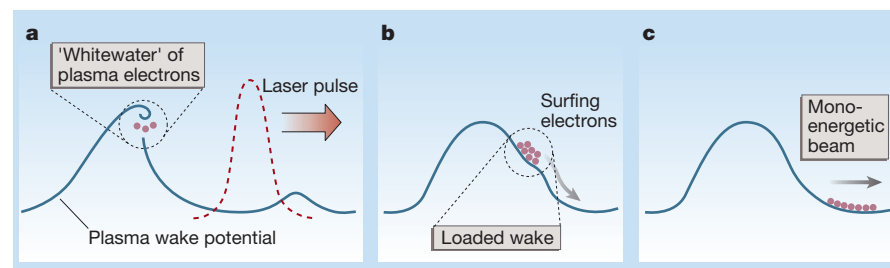


Figure 1 Wakefield acceleration. **a**, In a plasma excited by a laser pulse, the wake potential rises until it steepens and breaks. Electrons from the plasma are caught in the 'whitewater' and surf the wave. **b**, The load of the electrons deforms the wake, stopping further trapping of electrons from the plasma. **c**, As the electrons surf to the bottom of the wake potential, they each arrive bearing a similar amount of energy.

peak was expected to be a challenge for the field for several years to come.

Instead, the three groups reporting in this issue^{1–3} have found a new physical regime, in which electrons are 'self-injected' in a narrow region of space and made to surf as a single group, all reaching the same energy (Fig. 1). The three experiments are similar in many ways. In each of them, 10–30 terrawatts of laser power, in pulses 30–55 femtoseconds long, is focused into an ionized jet of gas roughly 2 mm long and with a particle density of $2 \times 10^{19} \text{ cm}^{-3}$; a nearly monoenergetic distribution of electrons is observed, with instrument-limited energy spreads of 2–24% at roughly 80–170 MeV. With up to a few times 10^9 electrons per beam, the energy densities in these experiments are a hundred to a thousand times higher than has previously been achieved. The angular spread of the beams is also about ten times tighter than before — comparable to the best of the beams produced by radio-frequency systems. Moreover, the pulse lengths of the beams are about 10 femtoseconds (10^{-14} s), making them attractive as potential radiation sources for ultrafast time-resolved studies in biology and physics.

Despite the similarities between the three experiments, it is the differences that have helped to identify the mechanisms responsible for their success. The three groups used different approaches to control what turns out to be a key factor — the interaction length in the plasma. The interaction length is the distance over which the particles surf the wake, and it is determined by either the end of the plasma or the weakening of the laser pulse through diffraction (the natural tendency of tightly focused light to spread). Geddes *et al.*¹ used a pre-formed plasma channel to guide the laser over several times the length that it would travel without diffraction in a vacuum; the groups of Mangles² and Faure³ used a larger laser spot size (up to 24 micrometres) to increase the interaction length. The groups describe essentially the same physics: first, the laser pulse evolves to become shorter and narrower; this creates a large wake that

traps electrons from the plasma; the loading of the wake with trapped particles turns off further trapping; and finally, 'dephasing' of the electrons as they outrun the wake creates a monoenergetic beam (basically, like marbles that roll to the bottom of a hill, they arrive at different times but end up at the same energy; Fig. 1).

Geddes *et al.*¹ emphasize the need for large interaction lengths to enable the electrons to dephase from the wave; their demonstration of guiding an intense laser in a plasma channel suggests a means of extending future wakefield accelerators beyond the millimetre scale. Mangles *et al.*², however, stress the need to reduce the interaction length to prevent the dephasing from becoming complete (the marbles reach the next hill and begin to slow down). Thus, as in the children's story *Goldilocks and the Three*

Bears, the interaction length must be not too long, nor too short, but just right.

There is still a long way to go from these experiments in the 100-MeV range to the frontiers of high-energy physics (it's likely that considerably more than 100,000 MeV needs to be available in a particle collision to produce a Higgs boson). The shot-to-shot stability and efficiency of these schemes also need to be improved. Nevertheless, these results represent the most significant step so far for laser-based accelerators, and should stimulate further advances in the near future. In particular, developments in high-power laser technology and plasma-channel production (particularly lower-density channels to increase the wake speed and hence the dephasing length) could both lead to the generation of beams of up to a few thousand MeV from a single-stage table-top device. Such accelerators would not only be more compact but would also exceed conventional sources in peak current, brightness and shortness of pulse duration. Wakefield acceleration may one day change the way we approach the physics and applications of particle beams.

Thomas Katsouleas is in the School of Engineering, University of Southern California, 3737 Watt Way, Los Angeles, California 90089-0271, USA.
e-mail: katsoule@usc.edu

1. Geddes, C. G. R. *et al.* *Nature* **431**, 538–541 (2004).
2. Mangles, S. P. D. *et al.* *Nature* **431**, 535–538 (2004).
3. Faure, J. *et al.* *Nature* **431**, 541–544 (2004).
4. Joshi, C. & Katsouleas, T. *Physics Today* **56**, No. 6, 47–51 (2003).
5. Modena, A. *et al.* *Nature* **377**, 606–608 (1995).
6. Malka, V. *et al.* *Science* **298**, 1596–1600 (2002).

Plant disease

Underground life for rice foe

Barbara Valent

We still have much to learn about the world's chief disease of rice — rice blast. That's clear from the finding that the culprit not only infects aerial plant tissues but can also invade roots like a typical root pathogen.

The diverse fungi that threaten the world's food crops are generally divided into those that infect plant structures above the ground and those that infect roots. Fungi that attack aerial plant structures use a few characteristic developmental pathways, and root-invading fungi — including symbiotic species that can be beneficial to plants — use different developmental routes. The rice blast fungus, which causes an annual loss of hundreds of millions of tonnes of rice worldwide, has become a model system for studying the aerial attack pathway. But, in a ground-breaking report that bridges the divide between the pathogenic lifestyles, Sesma and Osbourn¹ show that the foliar blast pathogen also invades roots, using a typical root-specific pathway (page 582 of this issue).

The status of rice blast as a model system for studying aerial plant infection is based on its continuing impact on world food production, its amenability to molecular and genetic analyses, and the well-defined developmental pathways it uses to invade aerial rice tissues. Infection occurs when airborne spores land on rice plants, sense the waxy aerial plant surface, and develop a dome-shaped organ called an appressorium. This produces phenomenal pressures — equivalent to those experienced in a deep-sea dive to 2,500 feet — enabling it to push a penetration peg through the tough surface layers that protect the plant². For the next week, the fungus spreads within the plant tissue and forms eyespot-shaped lesions (Fig. 1a), producing thousands of spores daily to invade new tissue.



Figure 1 Fungal targets. a, The rice blast fungus was thought to mount a direct attack only on above-ground plant tissues, such as the rice plant's grain-producing panicles (symptoms shown in main picture) and leaves (inset). b, The blast fungus's closest relatives block grain production by destroying plant roots; the main picture shows the 'whitehead' symptoms of one such fungus, that causing take-all root rot of wheat, and the inset shows the effects on a single wheat seedling. Sesma and Osbourn¹ have found that the rice blast fungus can in fact adopt both strategies. (Images in b courtesy A. Osbourn.)

It has generally been thought that this is the sole way in which rice blast infects undamaged rice plants. When Sesma and Osbourn¹ challenged roots with the blast fungus, however, it unexpectedly behaved like other root pathogens, using common tactics of root-invading fungi (see Fig. 1 on page 583). It formed large, highly pigmented runner hyphae — generated by the fusion of several of the fungus's long, thread-like cells — on the root surface. It formed infection pads typical of those used by root pathogens to gain entry to the root interior. And it produced two types of typical 'resting' structure, one inside the diseased root and a different one on the root surface. The authors also report that the blast fungus invades the root vascular system — the plant's pipeline for moving water and nutrients around — and then grows up this pipeline to produce lesions on aerial plant parts. This is a rather long list of developmental processes that no one realized the blast fungus could undergo.

Sesma and Osbourn support their microscopy studies by identifying blast genes specific for the aerial versus root attack strategies. They find that a gene essential for relaying the signal that the fungus is on a leaf is not required for root infection, and neither is a gene required for building up the high pressures that power foliar penetration. On the other side of the coin, by scanning the blast genome sequence³ the authors identified a gene similar to one required by another fungus to infect roots. Eliminating this gene (by genetic 'transformation') greatly reduced the blast fungus's ability to infect roots, but had no impact on leaf infection. It would be very unusual for an organism to retain genes for developmental processes it does not use, suggesting that the blast fungus at least occasionally infects roots in nature.

In one respect, it may not be so surprising that the blast fungus (*Magnaporthe grisea*) can behave as a classical root pathogen. Its closest relatives in the family Magnaporthaceae are all root pathogens, including the fungus that causes a major wheat disease called take-all⁴ (Fig. 1b). But despite decades of studies of rice blast, the possibility that it might infect roots had not received attention. This capability now joins other biological processes that have been observed in the laboratory but not in the field. For example, some rice blast strains produce sexual spores in the lab, but these have not been seen in the field. Curiously, these same strains also produce a type of asexual spore⁵ that is similar to those of the take-all fungus, raising the question of whether these 'microspores' — which are very different from the aerial asexual spores that cause foliar disease — might participate in root infection. Another fundamental question is whether the resting root structures are a factor in the over-winter survival of the blast fungus (as none of the three blast spore forms seems to be designed for long-term survival).

Compared with foliar diseases, much less is known about how root-infecting fungi cause disease, in part because soil fungi are difficult subjects for genetic transformation⁴. With the findings of Sesma and Osbourn¹, the genetically tractable blast fungus becomes a valuable resource for studying both of the pathogenic lifestyles encoded within its 11,000 or so genes. Understanding its evolutionary history — whether underground or aerial strategies came first, for example — will help us to assess the future evolutionary potential of numerous important cereal pathogens.

Compared with natural foliar resistance mechanisms, there is also much less known about resistance mechanisms in roots. For

example, aerial rice tissues express an arsenal of resistance genes, whose products recognize specific variants of the blast fungus and initiate defence responses⁶. But plant breeders have not been able to identify similar specific resistance to take-all in cereal roots⁴. Interestingly, however, Sesma and Osbourn show that specific resistance to rice blast in leaves also works against the blast fungus in roots, suggesting that this resistance could have significance in the field. Perhaps ongoing research into how to improve the efficacy and durability of foliar resistance genes (currently, the blast fungus quickly learns to avoid recognition through resistance genes that breeders introduce into field crops⁶) might some day help to control root disease, too.

What matters is developing effective methods to protect the world's food supply in an environmentally sustainable way. Rice blast alone is a major threat to world food security, and intensive farming practices such as a greater use of fertilizers will only increase its occurrence. But blast fungus threatens not only rice: it occurs as a series of host-specific forms that collectively infect many grasses, including the other major cereals. Blast emerged as a significant disease on wheat in Brazil in 1985 (ref. 7), and continues to affect production there. Wheat blast has not spread yet to other wheat-growing regions, but the threat remains.

Plant diseases will remain a moving target, as pathogens evolve new capabilities or move into niches released by the control of a competitor pathogen. Developing and deploying control strategies typically takes many years, so we must quickly understand the full biological potential of our enemies. We do not know whether root infection sometimes contributes to blast disease in the field, but we need to find out. Rice that is grown in flooded fields should not be vulnerable. But root infection could be a factor in the significant fraction of the world's rice crop that is not grown under flood, or in other cereal crops. When we are better able to control leaf blast disease, will the fungus switch to its underground strategy? One thing is clear: these are all important questions that wouldn't have been asked without the findings now reported¹.

Barbara Valent is in the Department of Plant Pathology, Kansas State University, 4024 Throckmorton Plant Sciences Center, Manhattan, Kansas 66506-5502, USA.
e-mail: bvalent@ksu.edu

1. Sesma, A. & Osbourn, A. E. *Nature* **431**, 582–586 (2004).
2. Talbot, N. J. *Annu. Rev. Microbiol.* **57**, 177–202 (2003).
3. International Rice Blast Consortium; www.riceblast.org
4. Freeman, J. & Ward, E. *Mol. Plant Pathol.* **5**, 235–252 (2004).
5. Kato, H. et al. *Ann. Phytopathol. Soc. Jpn* **60**, 175–185 (1994).
6. Zeigler, R. S., Leong, S. A. & Teng, P. S. (eds) *Rice Blast Disease* (CAB International, Wallingford, 1994).
7. Urashima, A. S., Igarashi, S. & Kato, H. *Plant Dis.* **77**, 1211–1216 (1993).

Human evolution

Pedigrees for all humanity

Jotun Hein

Simulations based on a model of human population history and geography find that an individual that is the genealogical ancestor of all living humans existed just a few thousand years ago.

Writing on page 562 of this issue, Rohde, Olson and Chang¹ address a simple but fascinating question: how far back in time must we go to find an individual who was the ancestor of all present-day humans? After a little consideration, the existence of such an individual (the 'universal ancestor' or, as the authors put it, our 'most recent common ancestor') should not surprise: I have two parents, four grandparents, and the growth in the population of my ancestors is close to exponential as I trace them back in time. This is true for anybody's ancestors, and there must soon be an overlap between the ancestors of two or more randomly chosen individuals (Fig. 1).

In simplified models, which assume random mating, the average number of generations back to a universal common ancestor has been estimated^{2–4} to be around $\log_2 n$, where n is the population size. So if, for instance, the present-day population were to consist of 1,000 people, the average number of generations back to the universal ancestor would be $\log_2(1,000)$ — about 10 generations. For populations of size 10^6 , or the present human population of size 6×10^9 , it would be 20 or 33 generations, corresponding to 500 or a bit more than 800 years, respectively (assuming a generation time of 25 years). This is surprisingly recent.

And an even more surprising conclusion from such models is that, only a little farther back in time, a large fraction of the population will be the ancestors of everybody alive today. The remaining individuals back then will be the ancestors of no one. As Rohde *et al.*¹ describe it, "When genealogical ancestry is traced back beyond the [universal ancestor], more and more people in earlier generations become ancestors of the [whole] present-day population". At a certain point in history (the 'identical ancestors' point), people can be divided into two groups: either they are common ancestors of all present-day humans, or their lineages have died out. Being the ancestor of only some living humans is not an option. At this point, Rohde *et al.* say, "everyone alive now had exactly the same ancestors". In the simplest model, the fraction of 'ancestors-of-all' is about 80%, and in most estimates so far, the time back to the 'identical ancestors' point is a bit less than twice the number of generations back to the first universal ancestor.

These estimates are not only astonishing, however; they are also unrealistically low,

because of the simplicity of the underlying models. Key missing factors are geography (which influences population structure) and history (which affects population growth), and these are the ingredients that Rohde *et al.* have taken seriously to arrive at more credible estimates of the time back to the universal and identical ancestors.

The authors carried out simulations based on several scenarios, incorporating different degrees of population growth and different degrees of isolation of subpopulations, with occasional migration linking these subpopulations. The authors' first model is relatively simple and includes up to ten large subpopulations, which exchange just one pair of migrants per generation. In one set of estimates based on this model, the mean time back to the universal ancestor is 2,300 years (76 generations, assuming a generation time of a bit less than 30 years) and to the identical ancestors it is 5,000 years (169 generations) — the time of Aristotle and the first pyramids, respectively. The latter date is especially startling: had you entered any village on Earth

in around 3,000 BC, the first person you would have met would probably have been your ancestor! A considerably more detailed model, which describes population density within continents, the opening of ports and more, does not change these estimates much.

The main weakness in the models comes from migration. As the authors point out, if one region is totally isolated (something that they do not simulate), with no migrants connecting it to other subpopulations, then the universal ancestor must logically have lived before the period of isolation began. Only after that period ends would the dates for the universal ancestor become less distant. Because of the effects of isolation, had we been living in 1700, say, and tried to work out when our universal and identical ancestors lived, the answers would have been further back in time than the answers we obtain now. Tasmania, for instance, was conceivably completely isolated at the time, and probably had been for millennia; this would therefore have pushed back the dates for universal and identical ancestry. So uncertainties about population structure introduce uncertainty into the proposed dates.

The genealogical questions addressed by Rohde *et al.* are distinct from questions about the history of our genetic material. In models that trace genetic material back in time, any given nucleotide position in our genomes can eventually be found in a single individual and on a single chromosome. Thus, being in the *pedigree* of all of humanity

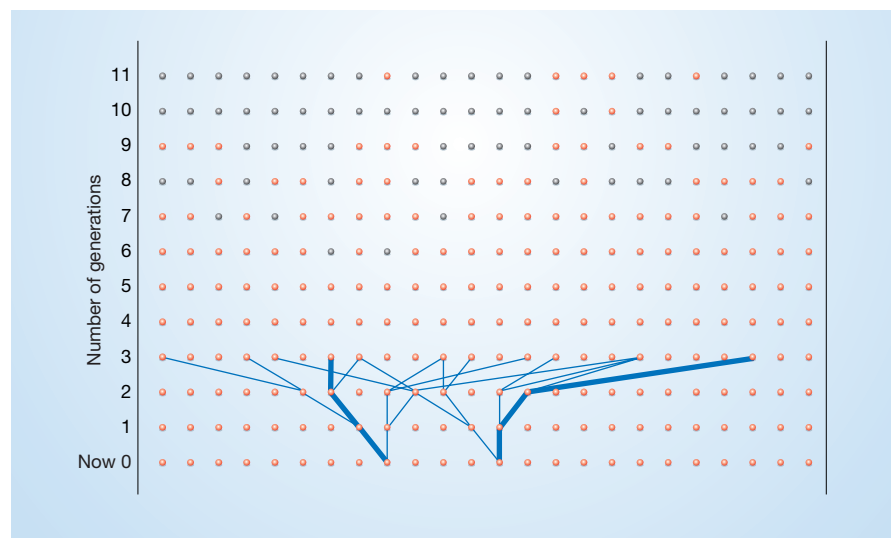


Figure 1 Searching for our universal common ancestor. The figure shows how the number of ancestors of two people alive today builds up in a manner that is close to exponential. Because the human population has a finite size, however, we do not need to go back many generations before we find an ancestor that is common to both people. The same applies in searching for the ancestor of all living humans (universal ancestors are represented as grey balls). In simplified models, the expected time back to this universal ancestor is $\log_2 n$, where n is the population size. If we were to trace not both parents of each individual, but only one random parent for each (thick lines), we would in effect be tracing the history of gene variants (alleles). In standard models, the number of generations back to the common ancestor of a particular allele will be of the order $2n$, which is much longer ago. If we trace the history of genomes, not genes, recombination would complicate matters; this genetic 'shuffling' ensures that each child does not inherit exactly the same genomic information as its siblings, and means that the genealogical relationship of different genome segments can be different.

Global change

Glacial pace picks up

When a huge chunk of Antarctic ice shelf broke up in 2002, it provided dramatic pictures (see right) for the world's press and a control experiment for researchers. The ice shelf, Larsen B, is a floating extension of the ice of the Antarctic peninsula. The collapse of a substantial part of it — more than 3,000 km² — was attributed to increasing temperatures and released shoals of icebergs into the Weddell Sea. But a southerly remnant remained in place, enabling ideas to be tested about how ice

shelves might affect glacier flow from the continental interior.

Two groups now report their results of satellite-tracking glacier behaviour in the region (E. Rignot *et al.* and T. A. Scambos *et al.* *Geophys. Res. Lett.* 10.1029/2004GL020697; 10.1029/2004GL020670). They found that five glaciers flowing into the area formerly buttressed by the ice shelf all accelerated at various times, whereas two farther south, which ran into the remnant ice shelf, did not. Speed of glacier flow is also

reflected in their thickness: higher flow rates stretch and thin the ice, in these cases yielding estimated rates of thinning of tens of metres per year.

The main implication is that ice shelves act as a restraint on glacier flow. This conclusion was by no means obvious. Earlier, theoretical studies gave conflicting results; and there are also possible confounding factors, such as water, produced by seasonal melting of surface ice, acting as a lubricant at the glacier base.

A prospect for the future — and



a worrying one as far as larger ice shelves and glaciers are concerned — is that a feedback system could kick in, accelerating glacier melting and producing significant rises in sea level.

Tim Lincoln

NASA/GSFC/LARC/JPL, MISR TEAM

does not imply that an individual makes a significant *genetic* contribution to the present population. In fact, that individual might have contributed nothing. This distinction is also illustrated by 'mitochondrial Eve' — the woman who purportedly lived hundreds of thousands of years ago and carried mitochondrial genes that are ancestral to all present mitochondrial genes. In Fig. 1 you would reach this Eve by tracing only female lineages backwards (rather than both lineages).

Universal common ancestry (in the pedigree sense) and genetic common ancestry thus occur on different timescales. The former is proportional to $\log_2 n$, and if you were to double the current population size, the expected time back to the universal ancestor would move back by only one generation in the simple model. But the time back to the genetic common ancestor is typically proportional to the population size, and so doubling the population size would double the time back to that kind of ancestor. The fact that the number of ancestors in a pedigree increases exponentially, whereas the number of genetic ancestors increases much more slowly, has the consequence that not many generations ago (about six), members of our pedigree existed that did not contribute to us genetically. So being somebody's great-great-great-grandparent is no guarantee of genetic relatedness. To properly understand genetic ancestry, we need the concept of the ancestral recombination graph^{5,6} — a generalization of traditional phylogeny that traces genetic material back in time in the presence of genetic recombination.

The increased ease of obtaining genome-sequence data from individuals, and the number of large-scale projects cataloguing variation in the human population, will increase our ability to test hypotheses about human history. Combining pedigree and genetic ancestry will become more and more important, both for data analysis and in

exploring properties of population models⁷. Many interesting questions lie ahead. For instance, how much genetic material (if any) did the universal ancestor pass on to the present population? What about that for a non-universal ancestor from the same time? In the idealized models, how far back would one have to go to find a single couple who are the lone ancestors of everybody? And how much could be known about humanity's pedigree if we knew the genome of everybody? ■

Jotun Hein is in the Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK.

e-mail: hein@stats.ox.ac.uk

1. Rohde, D. L. T., Olson, S. & Chang, J. T. *Nature* **431**, 562–566 (2004).
2. Kammerle, K. J. *Appl. Prob.* **27**, 880–885 (1989).
3. Chang, J. *Adv. Appl. Prob.* **31**, 1002–1026 (1999).
4. Derrida, B., Manrubia, S. C. & Zanette, D. H. *J. Theor. Biol.* **203**, 303–315 (2000).
5. Griffiths, R. C. *Theor. Popul. Biol.* **19**, 169–186 (1981).
6. Hudson, R. R. *Theor. Popul. Biol.* **23**, 183–201 (1983).
7. Hein, J. J., Schierup, M. H. & Wiuf, C. H. *Gene Genealogies, Variation and Evolution* (Oxford Univ. Press, 2004).

Cosmology

What is dark energy?

Lawrence M. Krauss

It seems that the rate of expansion of the Universe is accelerating, driven by the so-called dark energy. Is Einstein's cosmological constant behind it? There might be a way to find out.

The nature of the 'dark energy' that is causing the apparent accelerated expansion of the Universe is, without doubt, the biggest mystery in physics and astronomy. Although it was astrophysical observations of the acceleration that led to the discovery of dark energy, there are precious few tests that can be performed to work out what dark energy is — whether it is simply the rebirth of Einstein's cosmological constant, or whether it might stem from something even weirder. All the evidence so far is consistent with the existence of a cosmological constant, which, in modern language, is understood to be the quantum-mechanical energy associated with otherwise empty space. In *Physical Review D*, Kunz *et al.*¹ suggest, however, that by comparing data on a range of astrophysical phenomena, it might be possible to rule out a cosmological constant as the origin of dark energy.

Dark energy is perplexing. Physical theory

currently has no explanation of why the energy of empty space should be precisely zero (quantum-mechanical effects combined with relativity in fact predict quite the opposite). But it also gives no explanation of why that energy should not instead be so huge that it would dwarf all of the energy in anything else (making galaxy formation impossible). Yet arguments based on a host of different cosmological observations — even before the direct observation of the accelerated expansion — implied that the energy in empty space could not be more than three to four times greater than the energy contained in the matter and radiation of the Universe. To decide on what physics might be associated with dark energy, we have to rely on experiments and observations. No laboratory experiment we can imagine would be sensitive enough to do the job, so we are left with astrophysical probes. Which is where Kunz *et al.*¹ come in.

They propose a three-way comparison of data: of the expansion rate of the Universe as it changes with distance (from measurements made using type-Ia supernovae, which originally led to the discovery of dark energy^{2,3}); with measurements⁴ of the temperature fluctuations in the cosmic microwave background (the relic radiation of the Big Bang); and with measurements of the clustering of galaxies on large scales. Studies of the cosmic microwave background (CMB) have provided remarkably precise constraints on most major cosmological parameters, and are in some sense complementary to the limits derived using type-Ia supernovae. To describe the different possibilities for dark energy, an 'equation-of-state' parameter, w , is defined. This is the ratio of the pressure to the energy of the material. For the cosmological constant, w is exactly -1 ; any measured difference from this value would signal the need for another explanation. Data from the CMB, in combination with those from supernovae, currently limit w to the range $-1.2 < w < -0.8$, consistent with the value for a cosmological constant^{4,5}. (For comparison, w for matter is 0, and for radiation it is $1/3$.)

But Kunz *et al.*¹ point out that allowance should be made for a possible dynamical variation of w over time. The key new ingredient they throw into the mix is a comparison between the observed clustering of matter on large scales across the Universe and the predicted level of such clustering based on observations of the fluctuations of the CMB. It turns out that, because of the way that the dark energy comes to dominate the expansion of the Universe, the CMB temperature fluctuations should change on the largest angular scales (spanning more than about ten degrees across the sky) in a way that is sensitive to the dark-energy equation of state.

Now, from the CMB fluctuations on large scales, the overall scale of the clustering of matter in today's Universe — on the scale of galaxy clusters, millions of light years across — can be predicted: in the case that $w < -1$, the prediction is that clustering would be decreased. Thus, by comparing this prediction with measurements of galaxy clustering from large-scale redshift surveys, it might turn out that the value of w is not -1 — and so dark energy does not arise through a cosmological constant. The simplest interpretation of existing data suggests that this is not the case. But Kunz *et al.* point out that, first, there is a large spread in the data and, second, interpretation of the data is implicitly sensitive to assumptions about the nature of the dark energy. It is still possible that future studies could favour a value of w that is not -1 .

All of this points to what could be a big problem in cosmology lurking on the horizon. At present, the data are completely con-

sistent with a cosmological constant being behind dark energy. Unfortunately, however, there are other possible sources of dark energy — some of which I consider to be the best-motivated alternatives to a cosmological constant — that would produce a value for w of roughly -1 . Thus, measuring $w = -1$ does not uniquely specify the origin of dark energy. Only if w is not equal to -1 would we at least be able to say definitively that the dark energy is not associated with the ground-state quantum-mechanical energy of the vacuum.

Thus, some of us wake up in the middle of the night worrying that the discovery of dark energy may put cosmology on the same footing as particle physics, with all of the data that have come in over the years pointing consistently to exactly the same set of cosmic parameters, but without revealing any smoking-guns that could direct us to a fundamental theoretical rationale for why

the data take these values. I have even made a bet with physicists Stephen Hawking and Frank Wilczek that this will happen (then, even if my worst nightmare turns out to be true, I will at least get a few bottles of wine out of the bargain). On the other hand, perhaps the cross-comparison of present and future cosmological observations, along the lines proposed by Kunz *et al.*¹, will yield some new handle on this slippery problem. In that case, I might lose my bet, but the 'golden age' of cosmology would persist. ■

Lawrence M. Krauss is in the Department of Physics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106-7079, USA.

e-mail: krauss@genesis1.phys.cwru.edu

1. Kunz, M., Corasaniti, P.-S., Parkinson, D. & Copeland, E. J. *Phys. Rev. D* **70**, 041301 (2004).
2. Schmidt, B. P. *et al. Astrophys. J.* **507**, 46–63 (1998).
3. Perlmutter, S. J. *et al. Astrophys. J.* **517**, 565–586 (1999).
4. Spergel, D. N. *et al. Astrophys. J. Suppl. Ser.* **148**, 175–194 (2003).
5. Krauss, L. M. *Astrophys. J.* **604**, 481–483 (2004).

Cell biology

Sight at the end of the tunnel

Arthur Horwich

A chaperone molecule called trigger factor binds new polypeptide chains as they emerge from the protein-synthesis machinery. Crystal structures suggest that this molecule forms a hydrophobic 'cradle'.

Cells seem to leave nothing to chance, including the final step of information transfer — the folding of a newly made chain of amino acids into a three-dimensional, active, 'native' protein. Specialized proteins called molecular chaperones ensure that the process of folding, determined by the amino-acid sequence of a polypeptide chain, does not go awry^{1,2}. On page 590 of this issue, Ferbitz *et al.*³ present crystallographic images of a bacterial chaperone called trigger factor. The images provide clues to how this molecule interacts with the newly synthesized polypeptide chain as it emerges from a tunnel in the protein-synthesizing machinery (the ribosome), potentially cradling and protecting segments of the polypeptide.

Chaperones typically assist the folding process by specifically binding to polypeptides through a feature that is unique to non-native proteins — exposed hydrophobic surfaces. These surfaces become buried in the interior of a protein in its final form. Such hydrophobic regions, left to their own devices, can bind to each other, producing aggregates, which are not only a dead-end for protein function but also potentially toxic to the cell; for example, aggregates are found in several neurodegenerative diseases. Chaperones intervene by binding these exposed surfaces through a hydrophobic site of their own, preventing

aggregation and enabling productive folding when the chaperoned protein is released.

The long-awaited structure of the trigger-factor chaperone, presented by Ferbitz *et al.*³, reveals an extended arrangement of three domains — a 'crouching dragon' with a head, tail and arms — and a notable hydrophobic surface in the shape of a cradle that is exposed in the hollow between the tail and arms. Excitingly, Ferbitz *et al.* place this in a functional context by means of a second structure. This structure shows the tail portion of trigger factor in complex with the large subunit of the ribosome, suggesting the position of intact trigger factor as it might interact with the ribosome.

This second structure is a considerable technical achievement, involving astute evolutionary considerations, incisive biochemical analysis and some deft crystallography. The only ribosomal large subunit that has been observed at high resolution by X-ray crystallography is that from the archaeon *Haloarcula marismortui*. The structure of this subunit, presented several years ago⁴, provided unprecedented resolution of such features as the reaction centre, where peptide bonds are formed, and the exit tunnel. But archaea lack trigger factor, instead using other molecules to protect nascent chains.

The investigators had previously identified⁵ the contact site for trigger factor on the ribosome of the bacterium *Escherichia coli*,

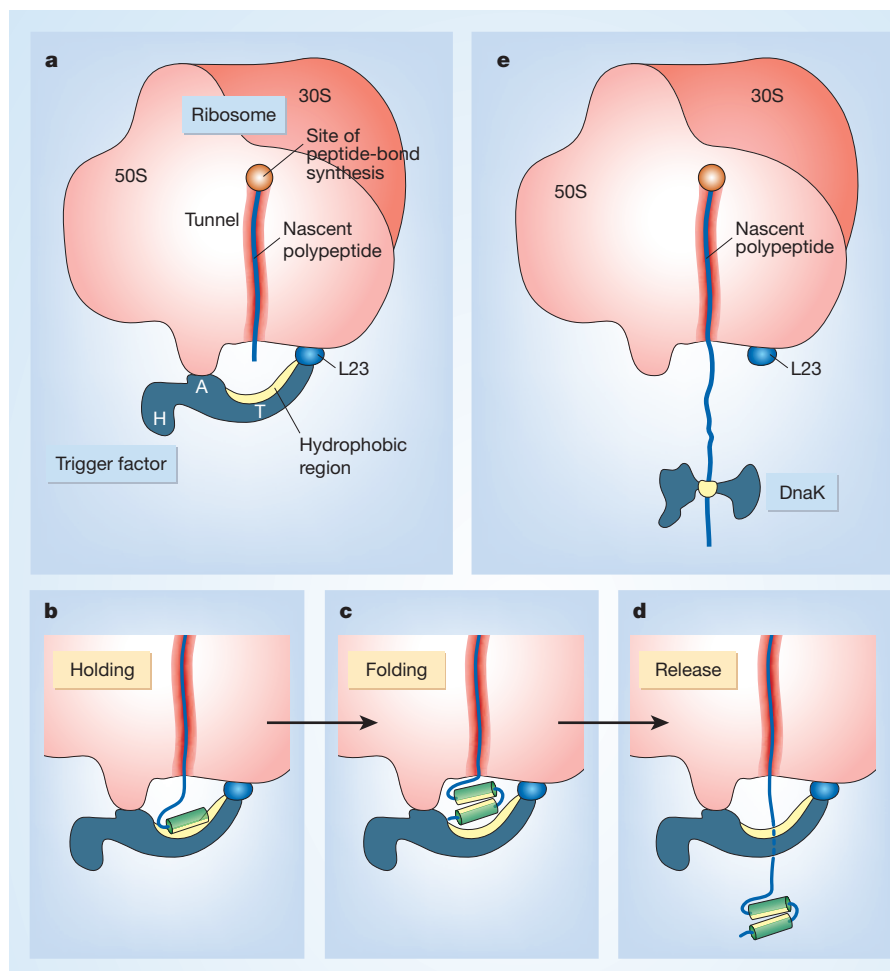


Figure 1 Chaperones at the ribosome. **a**, Ferbitz *et al.*³ have determined the crystal structure of the trigger-factor tail domain (T) in complex with a ribosome large subunit (50S), and used this to position their crystallographic model of intact trigger factor, including the arm (A) and head (H) domains. Trigger factor forms a cavity, with a hydrophobic lining, at the exit of a tunnel in the ribosome from which newly made polypeptide chains emerge. Its main contacts are with the ribosomal protein L23. **b–d**, How the nascent polypeptide might behave. **b**, 'Holding', where hydrophobic sections (yellow) of the nascent structure (here on an α -helix, symbolized by a barrel) could be stabilized by binding to trigger factor's hydrophobic surface. **c**, 'Folding', where a further stretch of emerging chain docks with that of the first, forming a substructure within the cavity. **d**, 'Release', where the substructure dissociates from trigger factor, by an unknown mechanism, and a new segment of chain enters the cavity. **e**, The binding of nascent polypeptide chains by chaperone DnaK is entirely different, and involves a narrow hydrophobic arch in DnaK⁹ (only the peptide-binding domain of DnaK is shown). The nascent chain might also acquire α -helical structure within the ribosomal tunnel^{12,13}, but this is not shown here.

finding that ribosomal protein L23 is particularly important. They reasoned that, because this contact site is highly conserved across the evolutionary kingdoms, a cross-kingdom complex of archaeal ribosome subunit and bacterial trigger factor should be possible. And they were right. Remarkably, the ribosome-binding tail domain of trigger factor, comprising 144 amino acids (out of 432), can bind to and crystallize as a complex with the ribosomal subunit. This allowed Ferbitz and colleagues to resolve the structure of part of this domain, consisting of the 40 amino acids that lie nearest to the ribosomal subunit.

Gratifyingly, the contacts that trigger factor makes with the large subunit are right

next to the exit of the ribosomal tunnel, and are made principally with the L23 protein. The authors then used the architecture of this fragment of trigger factor to superpose the entire stand-alone trigger-factor structure in relation to the ribosome. Startlingly, the chaperone seems to be positioned over the ribosomal tunnel exit, with its hydrophobic cradle providing a cavity into which the nascent chain would emerge (Fig. 1a).

The idea is inescapable that this cavity could produce a unique environment that can both physically protect the emerging nascent chain from protease enzymes (which could readily cleave a weakly structured polypeptide chain) and allow the stabilization of emerging hydrophobic parts of the

nascent chain (Fig. 1b). Furthermore, the cavity is large enough ($30 \times 35 \times \sim 20$ Å) to allow segments of the emerging nascent chain to fold partially inside it (Fig. 1c). For example, a local region, such as a protein domain, could potentially fold into its native form inside this space. Indeed, Ferbitz *et al.* used computer docking to show that the native form of the small protein lysozyme could fit into the space.

This model is very appealing, but it also raises a number of questions, not least about whether the positioning of the entire stand-alone model of trigger factor in relation to the ribosome is valid. For instance, might there be a structural rearrangement as the trigger factor binds to the ribosome–nascent-chain complex? There is a striking precedent for such a rearrangement: the so-called signal-recognition particle also associates with the ribosome L23 region, and it bends hugely when it binds to a ribosome–nascent-chain complex⁶. Such a transition would not necessarily be expected for trigger factor on functional grounds, but electron micrographs of ribosome-bound trigger factor with and without a nascent chain present would address this.

How general might the proposed mechanism for trigger factor be? A chaperone known as DnaK can entirely replace trigger factor in supporting protein folding^{7,8}. But, rather than forming a hydrophobic cradle, DnaK instead binds hydrophobic stretches of nascent chains through a small hydrophobic archway of its own⁹ (Fig. 1e). This suggests that a trigger-factor-like cavity is not essential to sustain correct protein folding.

It also remains unclear how the elongating polypeptide leaves trigger factor (Fig. 1d). Kinetic studies¹⁰ have suggested that trigger factor binds to the ribosome long enough for an entire chain to be synthesized, although these studies were done in the absence of polypeptide. This would imply either that the nascent chain exits through a side passage (front or back in Fig. 1) or that trigger factor changes its conformation to allow the nascent chain out. Alternatively, the growing nascent polypeptide might itself affect the residence time of trigger factor at the ribosome. Although several studies have suggested that nascent chains stabilize complexes of trigger factor with ribosomes (see, for example, ref. 11), perhaps transient release of trigger factor from the ribosome does nonetheless occur, such that the polypeptide controls the opening of the cavity and its own release.

More general questions about chaperone mechanisms also remain. Why haven't chaperones that interact with nascent chains been highly conserved in evolution (trigger factor, for instance, is found only in bacteria)? Are there kingdom-specific requirements that are related to particular proteomes or to cellular growth conditions? Or are there simply



100 YEARS AGO

The inability of a large number of skilful experimental physicists to obtain any evidence whatever of the existence of the n -rays, and the continued publication of papers announcing new and still more remarkable properties of the rays, prompted me to pay a visit to one of the laboratories in which the apparently peculiar conditions necessary for the manifestation of this most elusive form of radiation appear to exist. I went, I must confess, in a doubting frame of mind, but with the hope that I might be convinced of the reality of the phenomena, the accounts of which have been read with so much scepticism... I am obliged to confess that I left the laboratory with a distinct feeling of depression, not only having failed to see a single experiment of a convincing nature, but with the almost certain conviction that all the changes in the luminosity or distinctness of sparks and phosphorescent screens (which furnish the only evidence of n -rays) are purely imaginary. It seems strange that after a year's work on the subject not a single experiment has been devised which can in any way convince a critical observer that the rays exist at all.

R. W. Wood

From *Nature* 29 September 1904.

50 YEARS AGO

Jean Piaget's reputation as a psychologist in Great Britain is largely based upon a series of books written during 1925–32 dealing with the development of thought, language and moral judgment in the child. But, as he himself points out, this work was merely a prolegomena to his later investigations extending from 1937 to the present day... But though these researches are both theoretically and experimentally an advance upon his earlier work, they have, however, had little effect on English psychological thought... This is probably due to Piaget's introduction of a new and complex terminology, his use of symbolic logic, and the fact that his most important work remains untranslated... The most interesting conclusion which emerges from this important series of experimental researches is that mathematical concepts in their psychological development are ultimately based upon simple logical notions. Indeed, it might be said, without undue exaggeration, that Piaget's psychological studies are the genetic counterpart of Russell and Whitehead's attempt in "Principia Mathematica" to put mathematics on to a logical basis.

From *Nature* 2 October 1954.

many ways of bringing about a low-affinity, temporary interaction with a nascent chain — many ways for a chaperone 'midwife' to hold the baby?

Arthur Horwich is in the Department of Genetics and the Howard Hughes Medical Institute, Boyer Center, Yale School of Medicine, New Haven, Connecticut 06510, USA, and The Scripps Research Institute, La Jolla, California, USA.

1. Bukau, B., Deuring, E., Pfund, C. & Craig, E. A. *Cell* **101**, 119–122 (2000).
2. Hartl, F. U. & Hayer-Hartl, M. *Science* **295**, 1852–1858 (2002).
3. Ferbitz, L. et al. *Nature* **431**, 590–596 (2004).

4. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. *Science* **289**, 905–920 (2000).
5. Kramer, G. et al. *Nature* **419**, 171–174 (2002).
6. Halic, M. et al. *Nature* **427**, 808–814 (2004).
7. Deuring, E., Schulze-Specking, A., Tomoyasu, T., Mogk, A. & Bukau, B. *Nature* **400**, 693–696 (1999).
8. Teter, S. A. et al. *Cell* **97**, 755–765 (1999).
9. Zhu, X. et al. *Science* **272**, 1606–1614 (1996).
10. Maier, R., Eckert, B., Scholz, C., Lilie, H. & Schmid, F. -X. *J. Mol. Biol.* **326**, 585–592 (2003).
11. Hestekamp, T., Hauser, S., Lütke, H. & Bukau, B. *Proc. Natl. Acad. Sci. USA* **93**, 4437–4441 (1996).
12. Mingarro, I. et al. *BMC Cell Biol.* **1**, 3 (2000).
13. Woolhead, C. A., McCormick, P. J. & Johnson, A. E. *Cell* **116**, 725–736 (2004).

Biogeochemistry

Early options in photosynthesis

Nicolas Beukes

Reconstruction of an ancient marine environment from 3,400-million-year-old rocks in South Africa strengthens the case for the existence of photosynthetic microbes at that time — but adds a fresh twist.

Back in 1987, publication¹ of analyses of ancient rocks in Western Australia provided some startling news — the claim, based on structures interpreted as microfossils, of the existence of life by the end of the early Archaean eon, 3,400 million years ago. Subsequent investigations², however, led to the suggestion that the abundant organic material found in various rocks of that age had not been generated biologically but rather by abiotic reactions in hydrothermal systems. So here were two competing views of the early Earth: one in which Earth was already inhabited by relatively complex microbes, such as cyanobacteria, that produced oxygen as a by-product of photosynthesis; and another in which the environment was dominated by hydrothermal vents and springs spewing prebiotic organic soup into an uninhabited ocean.

On page 549 of this issue, Tice and Lowe³ add a twist to this debate with data from the 3,416-million-year-old rocks of the Buck Reef Chert in South Africa. They provide convincing evidence that the organic matter preserved in these rocks is of biological, not hydrothermal, origin. But they do not return to the view of an early Archaean Earth inhabited by oxygen-producing cyanobacteria. Rather, their picture is one in which non-oxygen-producing (anoxygenic) photosynthetic microbes existed in an ecosystem that was fundamentally different from that of today.

Like the Western Australian material that is the subject of the earlier controversy, the Buck Reef rocks are composed of chert, a sedimentary rock made almost entirely out of microcrystalline quartz. The chert contains abundant organic inclusions that have been heated to such a degree that they contain no extractable biomolecules, but which retain spectacularly preserved structures from the time of their deposition. Tice and Lowe's

evidence that these carbonaceous inclusions are of biological origin comes partly from their morphology: some resemble microbial mats whereas others appear to be sand- and silt-sized grains formed by erosion of the mats.

However, the real robustness of their interpretations lies in their reconstruction of the environmental setting in which the Buck Reef Chert formed, and their ideas about how the distribution, morphology and structuring of the carbonaceous matter correlate with those settings. Tice and Lowe show that the Buck Reef Chert has three main components: a layer that was deposited in evaporative ponds behind an old shoreline; a carbon-rich, black-and-white-banded chert unit, deposited in a shallow nearshore environment that was occasionally stirred by storms and large waves; and a banded, iron-rich chert that formed offshore, below the base of storm waves at depths of more than 200 m.

From this reconstruction of sedimentary environments, the authors conclude that the mat-like organic laminations in black chert apparently had ecological control over their distribution. The laminations are only present in banded chert, deposited in a shallow marine environment, within the depth to which light could penetrate the water column (the euphotic zone). The distribution of these distinctive organic morphologies is best explained by their being of biological origin.

This result takes our understanding of early Archaean biota beyond the hydrothermal debate, and greatly improves the case for the existence of photosynthetic organisms in the early Archaean. Previous arguments for that rested primarily on interpretations of the morphologies of microfossils¹, and structures presumed to have been formed by cyanobacteria (stromatolites), and on the carbon isotopic composition of early organic

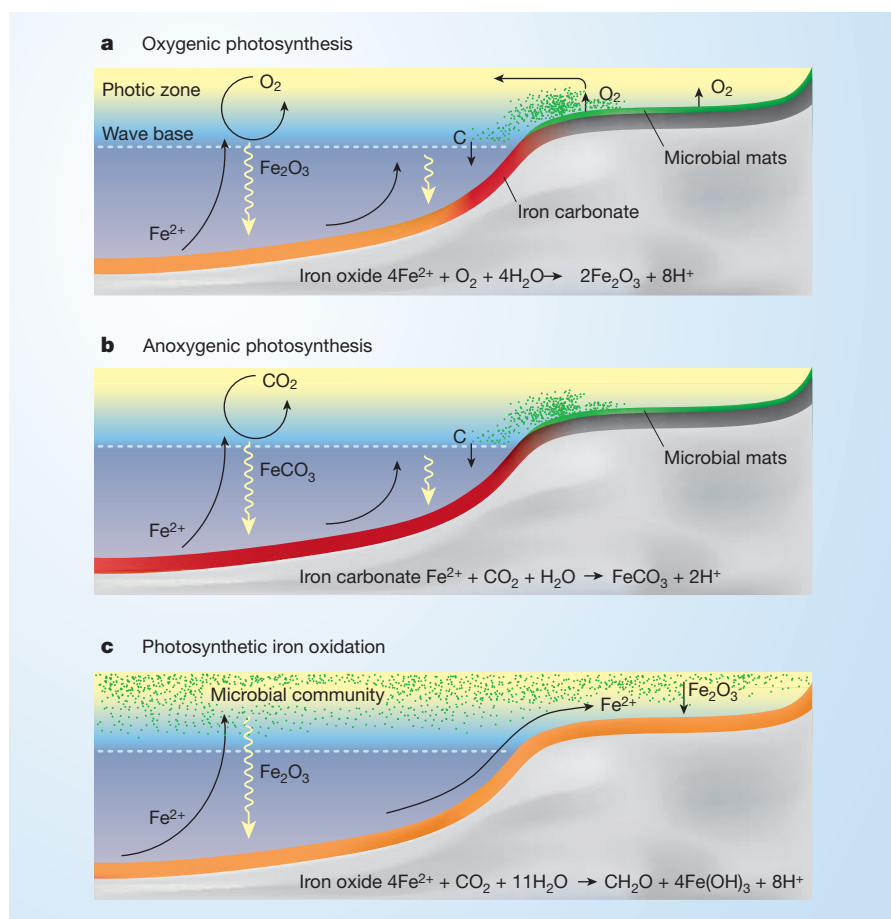


Figure 1 Contrasting models of iron-mineral deposition in the Archaean ocean. **a**, The classical model⁵, in which free oxygen, ultimately derived from oxygenic microbial photosynthesis, precipitates iron oxides from a stratified ocean. Iron carbonate (FeCO_3) can form in intermediate environments between the deep- and shallow-water settings. **b**, A model, based on Tice and Lowe's data³ from the Buck Reef Chert, of iron-carbonate formation under anaerobic conditions from a stratified ocean with a CO_2 -rich upper layer. Here, photosynthesis is anoxygenic and no iron oxides are precipitated. **c**, A recent model that requires dissolved ferrous iron to be present in the euphotic zone for oxidation by anaerobic photosynthetic bacteria⁶ in the open ocean. In **a** and **b**, deposition of the iron minerals is decoupled from primary biological productivity; in **c** it is directly coupled.

matter⁴. Both types of evidence were ambiguous for deposits of that age. Tice and Lowe's combination of data — carbon isotopic evidence and the apparent restriction of microbial mats to the euphotic zone — makes a much stronger case for photosynthesis as the primary mode of carbon fixation by microbial communities during deposition of the Buck Reef Chert.

But what kind of photosynthetic microbes were present? Here the iron-rich chert, representing the deep-water deposits of the Buck Reef Chert, is the key to Tice and Lowe's interpretation. It is composed of alternating bands of relatively pure white chert and chert containing fine laminations of iron carbonate (FeCO_3 , also known as siderite). It is analogous to the iron-carbonate formations commonly seen in the transition from black carbonaceous chert to iron-oxide (Fe_2O_3 and Fe_3O_4) deposits in many iron formations⁵. Models of iron deposition suggest that iron minerals periodically precipitated at the mixing interface of a deep, iron-rich

layer and a shallow, iron-depleted layer in a stratified Archaean ocean⁵. As Tice and Lowe point out, the abundance of iron carbonate in the deep-water deposits of the Buck Reef Chert, and its scarcity or absence in the shallow-water deposits, support this general concept of a stratified ocean.

Classic models of the deposition of iron minerals invoke the presence of free oxygen in shallow water to account for precipitation of iron oxide (Fig. 1a). In these models, iron carbonate is derived from reduction of iron oxide by organic matter washed in from shallow, nearshore environments. However, based on the observations that primary sedimentary iron oxides are absent from deep-water iron-rich chert, and that iron carbonate is found both in isolation and associated with carbonaceous matter in the Buck Reef Chert, Tice and Lowe³ propose that iron carbonate was directly precipitated from sea water. They thus favour the existence of an anaerobic environment in which anoxygenic photosynthetic microbes

inhabited the euphotic zone on a shallow platform (Fig. 1b). Iron carbonate precipitated deeper down, at the interface between a deep iron-rich layer and a shallow CO_2 -dominated layer in the Archaean ocean.

In recent years, a popular model for iron deposition during the Archaean has been one in which anaerobic photosynthetic bacteria oxidize ferrous iron in the euphotic zone, resulting in the precipitation of ferric iron oxides (Fig. 1c)⁶. Another possibility is that photochemical oxidation of ferrous iron under anaerobic conditions produced the same result⁷. But Tice and Lowe's reconstruction of the Buck Reef Chert shows that the entire euphotic zone of the stratified water column was depleted in iron. So it is unlikely that either of these two light-dependent mechanisms could have been in operation. Because they are the only known mechanisms that can oxidize ferrous to ferric iron under anaerobic conditions, it follows that free oxygen must have been available to produce the iron-oxide formations laid down during the Archaean. Many such formations are known in middle to early Archaean deposits, including that at Isua, Greenland⁸, which is about 3,800 million years old.

There are two known processes that could have produced free oxygen in Archaean times: light-mediated dissociation of water vapour in the upper atmosphere, and bacterial photosynthesis. Most models predict that oxygen production from water vapour would have been insignificant; it is also hard to imagine how that oxygen could have been transported through an otherwise reducing atmosphere and come into contact with the reduced iron dissolved in ocean water. Despite Tice and Lowe's conclusion that anoxygenic photosynthetic microbes were present during deposition of the Buck Reef Chert, does this mean that oxygenic photosynthesis developed very early on in Earth's history — perhaps even before deposition of the 3,800-year-old Isua formation?

This is a hypothesis that has been around for some time⁹, and Tice and Lowe provide a guide as to how to test it more rigorously in future. Before drawing conclusions about the global nature of environments or ecosystems early in the world's history, we need detailed field studies on iron formations and associated rocks in as many Archaean settings as possible to evaluate the geochemical and biological legacy left by local conditions. ■

Nicolas Beukes is in the Department of Geology, Rand Afrikaans University, Auckland Park 2006, South Africa.

e-mail: njb@rau.ac.za

- Schopf, J. W. & Packer, B. M. *Science* **237**, 70–73 (1987).
- Brasier, M. D. *et al.* *Nature* **416**, 76–81 (2002).
- Tice, M. M. & Lowe, D. R. *Nature* **431**, 549–552 (2004).
- Schidlowski, M. *Nature* **333**, 313–318 (1988).
- Klein, C. & Beukes, N. J. *Econ. Geol.* **84**, 1733–1774 (1989).
- Konhauser, K. O. *et al.* *Geology* **30**, 1079–1082 (2002).
- Cairns-Smith, A. G. *Nature* **276**, 807–808 (1978).
- Dymek, R. F. & Klein, C. *Precamb. Res.* **39**, 247–302 (1988).
- Cloud, P. E. *Am. Sci.* **62**, 389–390 (1974).

Anatomy

Elbow room for top predators

Zool. J. Linn. Soc. **142**, 91–104 (2004)

Large mammalian carnivores are often thought of as adopting one of two hunting styles: chasing down their prey, or overpowering them in an ambush. But does an animal's anatomy reveal which strategy it has adopted? It seems that it does — the forelimbs give the game away.

Ki Andersson measured the dimensions of forelimb joints (the equivalent of our elbow) in a range of mammal meat-eaters. Those that chase their prey in sustained pursuits, such as wolves and hyenas, have narrow forelimb joints that offer stability for running but restrict other movement. In contrast, species such as bears and the biggest big cats, which ambush and grapple prey, have wider elbow joints that give them dexterous paws.

These patterns are shaped by strong natural selection, Andersson argues. Cheetahs, perhaps the most noted exponent of the chase, have narrow joints indicative of their style, despite being closely related to other cats that wield their paws with the help of bulkier elbows.

Andersson has also studied the elbows of fossil mammals related to badgers and otters, and concludes that the technique could yield insights into the lifestyles of extinct species. For example, the African *Ekorus ekakeran* seems to have been a chaser, whereas *Megalictis ferox*, from North America, probably used an ambush strategy.

Michael Hopkin

Ceramics

A mark of imperfection

Chem. Mater. **16**, 3641–3646 (2004)

Japanese pottery from Bizen, decorated in the Hidasuki style (pictured), is made from yellowish clay streaked with fiery patterns in shades of red. No glaze is used; the clay vessel is simply bundled up in rice straw during firing in the kiln, and the straw somehow imparts the bright coloration.

The technique was discovered by chance more than 1,000 years ago, and is considered to imbue the stoneware with the quality of *wabi-sabi*: an earthy, impermanent and incomplete beauty. The colour is known to be caused by iron in the clay, which reacts with alkaline material in the straw. Yoshihiro Kusano *et al.* have now deciphered the microscopic chemistry of this process.

Clay from Bizen typically contains 2–3% by weight of iron oxide, and when heated, this combines with the aluminosilicate clay to form the pale mineral mullite, doped with iron. But potassium in rice straw induces instead the formation of corundum (aluminium oxide) crystals, on which smaller haematite particles stick like barnacles. This seems to be the colorant of the Hidasuki pattern; the precise hue depends on the size of the haematite crystals.

Philip Ball



Chemistry

Temperature-sensing trees

J. Am. Chem. Soc. doi:10.1021/ja047755g (2004)

Dendrimers are the 'spreading chestnut trees' of the molecular world. These spherical molecules branch out from a central core, ending in a cluster of atomic twigs.

Yasuhiro Haba *et al.* now report a temperature-sensitive dendrimer, made by hanging an isobutyramide chemical group from the end of each twig. These dendrimers are extremely soluble in phosphate-buffered water at room temperature. But as soon as the temperature rises above a critical temperature, the dendrimers quickly fall out of solution and the mixture becomes milky.

Interestingly, the authors find that increasing the number of branching points in the dendrimer lowers its critical temperature significantly: a dendrimer with five generations of branching becomes insoluble above 42 °C, but one with only three branching generations has a critical temperature of around 75 °C.

It had seemed that large differences in the molecular weight of a dendrimer only altered this critical temperature by a few degrees. Haba *et al.* believe that the effect they see might be caused by the marked increase in the density of the isobutyramide groups at the edges of the larger dendrimer.

Mark Peplow

Physics

Levitating femtodroplets

Appl. Phys. Lett. doi:10.1063/1.1781735 (2004).
Video footage at http://ftp.aip.org/epaps/appl_phys_lett/E-APPLAB-85-040432/

Shrinking a laboratory to the size of a silicon chip is a promising way to automate rapid chemical analyses of tiny amounts of material. Conventional microfluidic devices can move microlitre quantities of liquids around, but surface contact can slow down the samples or even degrade them.

I. F. Lyuksyutov *et al.* describe a diamagnetic levitation system that can process droplets that are a billion times smaller — femtodroplets. The tiny droplets can be positioned with 300-nm accuracy and merged with each other; solid particles can be precisely rotated and assembled into chains.

The system uses two oppositely magnetized neodymium–iron–boron permanent magnets that confine droplets in the space between them, while magnetic or electric fields from electrodes in the base are used to push the droplets back and forth. By trapping droplets between magnets, the authors effectively create miniature beakers for contained chemical reactions, and have tested their system with, for example, droplets containing single red blood cells and alcohol solutions.

Mark Peplow

Cancer

Hedgehog targeted

Cancer Cell **6**, 229–240 (2004)

Medulloblastomas are malignant brain tumours that occur in children. Existing treatments, which involve surgery, chemotherapy and radiation, can cause unacceptable side effects, so the hunt is on for an effective, non-toxic alternative. Justyna T. Romer *et al.* now report studies of one potential alternative in mice.

The authors' study used mice that had been genetically altered to develop medulloblastomas. They treated the animals with a small molecule that inhibits a particular cell signalling pathway — that involving the Sonic Hedgehog protein. When the drug was given at low doses, tumour cells proliferated less frequently and began to die off. Higher doses completely eradicated the tumours within two weeks, and helped the animals to stay tumour-free for longer periods than mice that were not treated with the drug.

Romer *et al.* hope that this or similar molecules will prove useful in treating medulloblastomas and other cancers in which the Sonic Hedgehog pathway has gone awry. But ultimately, they caution, many compounds, targeting many cell-growth pathways, may be needed.

Helen Pilcher

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

The Athens Olympic Games could be viewed as another giant experiment in human athletic achievement. Are women narrowing the gap with men, or falling further behind? Some argue that the gains made by women in running events between the 1930s and the 1980s are decreasing as the women's achievements plateau¹. Others contend that there is no evidence that athletes, male or female, are reaching the limits of their potential^{1,2}.

In a limited test, we plot the winning times of the men's and women's Olympic finals over the past 100 years (ref. 3; for data set, see supplementary information) against the competition date (Fig. 1). A range of curve-fitting procedures were tested (for methods, see supplementary information), but there was no evidence that the addition of extra parameters improved the model fit significantly from the simple linear relationships shown here. The remarkably strong linear trends that were first highlighted over ten years ago² persist for the Olympic 100-metre sprints. There is no indication that a plateau has been reached by either male or female athletes in the Olympic 100-metre sprint record.

Extrapolation of these trends to the 2008 Olympiad indicates that the women's 100-metre race could be won in a time of 10.57 ± 0.232 seconds and the men's event in 9.73 ± 0.144 seconds. Should these trends continue, the projections will intersect at the 2156 Olympics, when — for the first time ever — the winning women's 100-metre sprint time of 8.079 seconds will be lower than that of the men's winning time of 8.098 seconds (Fig. 1). The 95% confidence intervals, estimated through Markov chain Monte Carlo simulation⁴ (see supplementary information), indicate that this could occur as early as the 2064 or as late as the 2788 Games.

This simple analysis overlooks numerous confounding influences, such as timing accuracy, environmental variations, national boycotts and the use of legal and illegal stimulants. But it is also defended by the limited amount of variance that remains unexplained by these linear relationships.

So will these trends continue and can women really close the gap on men? Those who contend that the gender gap is widening

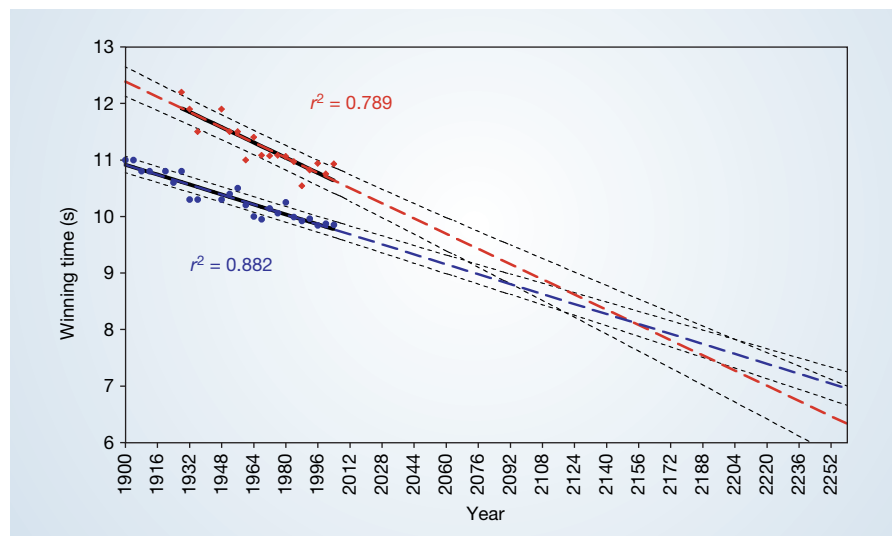


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

say that drug use explains why women's times were improving faster than men's, particularly as that improvement slowed after the introduction of drug testing¹. However, no evidence for this is found here. By contrast, those who maintain that there could be a continuing decrease in gender gap point out that only a minority of the world's female population has been given the opportunity to compete (O. Anderson, www.pponline.co.uk/encyc/0151.htm).

Whether these trends will continue at the Beijing Olympics in 2008 remains to be seen. Sports, biological and medical sciences should enable athletes to continue to improve on Olympic and world records, by fair means or foul⁵. But only time will tell whether in the 66th Olympiad the fastest human on the planet will be female.

Andrew J. Tatem*, **Carlos A. Guerra***, **Peter M. Atkinson†**, **Simon I. Hay*‡**

*TALA Research Group, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK
e-mail: andy.tatem@zoology.oxford.ac.uk

†School of Geography, University of Southampton, Highfield, Southampton SO17 1BJ, UK

‡Public Health Group, KEMRI/Wellcome Trust Research Laboratories, PO Box 43640, 00100 GPO, Nairobi, Kenya

1. Holden, C. *Science* **305**, 639–640 (2004).

2. Whipp, B. J. & Ward, S. A. *Nature* **355**, 25 (1992).

3. Rendell, M. (ed.) *The Olympics: Athens to Athens 1896–2004* 338–340 (Weidenfeld and Nicolson, London, 2003).

4. Gilks, W. R., Thomas, A. & Spiegelhalter, D. J. *Statistician* **43**, 169–178 (1994).

5. Vogel, G. *Science* **305**, 632–635 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Lung cancer

Intragenic ERBB2 kinase mutations in tumours

The protein-kinase family is the most frequently mutated gene family found in human cancer and faulty kinase enzymes are being investigated as promising targets for the design of antitumour therapies. We have sequenced the gene encoding the transmembrane protein tyrosine kinase ERBB2 (also known as HER2 or Neu) from 120 primary lung tumours and identified 4% that have mutations within the kinase domain; in the adenocarcinoma subtype of lung cancer, 10% of cases had mutations. ERBB2 inhibitors, which have so far proved to be ineffective in treating lung cancer, should now be clinically re-evaluated in the specific subset of patients with lung cancer whose tumours carry ERBB2 mutations.

The successful treatment of chronic myelogenous leukaemia with a drug (known as imatinib, marketed as Gleevec) that inhibits a mutant protein kinase has fostered interest in the development of other kinase inhibitors¹. Gefitinib, an inhibitor of the epidermal growth-factor receptor (EGFR), induces a marked response in a small subset of lung cancers; activating mutations have been found in the EGFR gene in tumours that respond to gefitinib but are rare in those that do not respond^{2,3}. The response to gefitinib as a treatment for lung cancer therefore seems to be predicated upon the presence of an EGFR mutation in the tumour.

ERBB2 and EGFR are both members of the EGFR kinase subfamily. Receptor oligomerization triggers signalling cascades implicated in cell growth, differentiation and survival. As part of an evaluation of these and other kinase genes for their involvement in human cancer and in order to find potential targets for mutant-kinase inhibitors, we sequenced the entire coding sequence and the exon/intron boundaries of the *ERBB2* gene in 120 primary lung tumours.

We identified three unambiguous somatic mutations (which were not present in normal DNA from the same individuals), two instances of an in-frame insertion (PD1353a and PD0258a) and a missense substitution (PD0270a) (Table 1). Two additional likely somatic mutations were found in tumours for which no normal tissue was available (one of these is a further instance of the previously observed in-frame insertion; the second is a different in-frame insertion, two amino acids distal to the other insertion). All mutations were located in the kinase domain. These in-frame insertions are adjacent to, and the missense mutation overlaps with, the analogous structural region of the in-frame *EGFR* deletions that are associated with some lung tumours^{2,3} (Fig. 1).

Immunocytochemical staining for ERBB2 revealed no differences between tumours with or without *ERBB2* mutations, indicating that overexpression probably does not accompany the mutation. *ERBB2* amplification was found in 1/49 adenocarcinomas and 1/14 large-cell carcinomas (neither of which had an intragenic mutation). None of the cancers associated with *ERBB2* mutation had mutations in *KRAS2*, *NRAS* or *BRAF*, genes that have also been implicated in the development of lung cancer⁴.

We determined the complete *ERBB2* coding sequence in 18 breast, 20 gastric and 15 testicular tumours; the kinase domain was sequenced in 303 primary cancers, including 31 colorectal, 40 renal, 27 ovarian, 10 glioma, 9 acute lymphocytic leukaemia, 20 myeloproliferative disease, 76 sarcoma, 11 papillary thyroid, 23 bladder, 56 additional breast and 235 cancer cell lines (see supplementary information). Three further somatic mutations were found, all in the kinase domain (Table 1); a mutation was

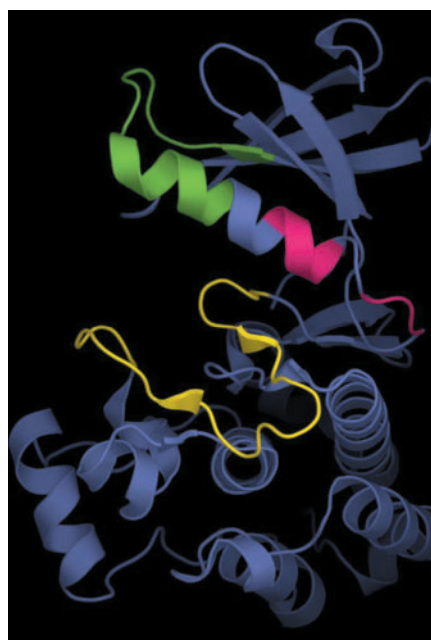


Figure 1 Similar positioning within the epidermal growth-factor receptor (EGFR) kinase domain (database accession numbers MMD:20494/PDB:1M17) of the EGFR and ERBB2 mutations that are found in a proportion of lung tumours. The composite position of reported EGFR deletions^{2,3} is indicated in green; the relative positions of the ERBB2 insertions described here are mapped onto the EGFR sequence and are shown in pink. The first third of the activation loop of the kinase domain is indicated in yellow for orientation.

also detected in a primary gastric cancer between two in-frame insertions.

In the lung tumours, all of the intragenic *ERBB2* mutations that we found were in adenocarcinomas (Table 1). The frequency was 4.2% (5/120) in non-small-cell lung carcinomas (NSCLCs) overall and 9.8% (5/51) in adenocarcinomas. By comparison, we found *EGFR* mutations in 2% (2/120) of NSCLCs and 4% (2/51) of adenocarcinomas, in agreement with a comparable series described previously³. None of these had an *ERBB2* mutation. Four out of five cases with *ERBB2* mutations were current or ex-smokers (*EGFR* mutation cases are predominantly found in never-smokers^{2,3}).

Although amplification of *ERBB2* has been demonstrated in 20% of breast cancers⁵ and occurs at a lower frequency in other cancers⁶, intragenic mutations in

ERBB2 in human cancer have not previously been reported. The pattern of *ERBB2* mutations, supported by precedents from other mutated kinases implicated in cancer development, strongly indicates that these mutations activate the ERBB2 kinase.

The drug trastuzumab (marketed as Herceptin), a humanized antibody against the extracellular domain of ERBB2, has been approved for treatment of metastatic breast cancer and is most effective in breast cancers with *ERBB2* amplification⁷. The presence of a mutation appears to be a major determinant of response to therapy, as is the case with gefitinib and the EGFR mutations^{2,3}. But results from phase II trials of trastuzumab as a treatment for NSCLC have not shown any advantage for most patients⁷ and have provided insufficient evidence to proceed to phase III trials⁸. However, our findings, coupled with results from gefitinib inhibition of EGFR mutants, indicate that targeting of ERBB2 with antibodies or small-molecule inhibitors should be considered in cases of lung adenocarcinoma that have demonstrable *ERBB2* mutations.

Cancer Genome Project and Collaborative Group*

e-mail: paf@sanger.ac.uk

*Philip Stephens¹, Chris Hunter¹, Graham Bignell¹, Sarah Edkins¹, Helen Davies¹, Jon Teague¹, Claire Stevens¹, Sarah O'Meara¹, Raffaella Smith¹, Adrian Parker¹, Andy Barthorpe¹, Matthew Blow¹, Lisa Brackenbury¹, Adam Butler¹, Oliver Clarke¹, Jennifer Cole¹, Ed Dicks¹, Angus Dike¹, Anja Drozd¹, Ken Edwards¹, Simon Forbes¹, Rebecca Foster¹, Kristian Gray¹, Chris Greenman¹, Kelly Halliday¹, Katy Hills¹, Vivienne Kosmidou¹, Richard Lugg¹, Andy Menzies¹, Janet Perry¹, Robert Petty¹, Keiran Raine¹, Lewis Ratford¹, Rebecca Shepherd¹, Alexandra Small¹, Yvonne Stephens¹, Calli Tofts¹, Jennifer Varian¹, Sofie West¹, Sara Widaa¹, Andrew Yates¹, Francis Brasseur¹, Colin S. Cooper¹, Adrienne M. Flanagan¹, Margaret Knowles¹, Suet Y. Leung¹, David N. Louis¹, Leendert H. J. Looijenga¹, Bruce Malkowicz², Marco A. Pierotti¹⁰, Bin Teh¹¹, Georgia Chenevix-Trench¹², Barbara L. Weber¹³, Siu T. Yuen¹⁴, Grace Harris¹⁴, Peter Goldstraw¹⁴, Andrew G. Nicholson¹⁴, P. Andrew Futreal¹, Richard Wooster¹, Michael R. Stratton^{1,3}

¹Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK; ²Ludwig Institute for Cancer Research, 1200 Brussels, Belgium; ³Institute of Cancer Research, Sutton SM2 5NG, UK;

⁴Institute of Orthopaedics, University College London, Stanmore HA7 4LP, UK; ⁵Cancer Research UK Clinical Centre, St James's University Hospital, Leeds LS9 7TF, UK; ⁶University of Hong Kong

Department of Pathology, Queen Mary Hospital, Hong Kong; ⁷Massachusetts General Hospital East, Molecular Pathology Unit, Charlestown, Massachusetts 02129, USA; ⁸Laboratory of

Pathology/Experimental Patho-Oncology, Erasmus MC University Medical Center Rotterdam, Daniel den Hoed Cancer Center, Josephine Nefkens Institute, 3000 DR Rotterdam, The Netherlands;

⁹University of Pennsylvania Division of Urology, Philadelphia, Pennsylvania 19104, USA; ¹⁰Instituto Nazionale Tumori and FIRC Institute of Molecular Oncology, Milan, Italy; ¹¹Laboratory of Cancer

Genetics, Van Andel Research Institute, Grand Rapids, Michigan 49503, USA; ¹²Queensland Institute of Medical Research, Royal Brisbane Hospital, Herston, Queensland 4029, Australia; ¹³University of Pennsylvania Cancer Centre, Philadelphia, Pennsylvania 19104-6160, USA; ¹⁴Royal Brompton Hospital, London SW3 6NP, UK.

1. Sawyers, C. L. *Genes Dev.* **17**, 2998–3010 (2002).
2. Paez, J. G. *et al. Science* **304**, 1497–1500 (2004).
3. Lynch, T. J. *et al. N. Engl. J. Med.* **350**, 2129–2139 (2004).
4. Futreal, P. A. *et al. Nature Rev. Cancer* **4**, 177–183 (2004).
5. Slamon, D. J. *et al. Science* **235**, 177–182 (1987).
6. Klapper, S. *et al. Adv. Cancer Res.* **1**, 25–79 (2000).
7. Hirsch, F. R. & Langer, C. J. *Semin. Oncol.* **31**, 75–82 (2004).
8. Rosell, R. J. *Clin. Oncol.* **22**, 1171–1173 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Table 1 ERBB2 mutations in primary tumours

Sample	Tumour/histology	Nucleotide*	Amino acid*
PD1353a	NSCLC adenocarcinoma	2322 ins/dup(GCATACGTGATG)	ins774(AYVM)
PD0258a	NSCLC adenocarcinoma	2322 ins/dup(GCATACGTGATG)	ins774(AYVM)
PD0317a	NSCLC adenocarcinoma	2322 ins/dup(GCATACGTGATG)	ins774(AYVM)
PD0319a	NSCLC adenocarcinoma	2335 ins(CTGTGGGCT)	ins779(VGS)
PD0270a	NSCLC adenocarcinoma	TT2263-4CC	L755P
PD1487a	Glioblastoma	G2740A	E914K
PD1403a	Gastric tumour	G2326A	G776S
PD0888a	Ovarian tumour	A2570G	N857S

NSCLC, non-small-cell lung carcinoma; ins, insertion; dup, duplication (see supplementary information); amino-acid residues are shown in the single-letter notation and substitutions are represented as wild-type residue/position/mutant residue.

*Numbering represents the position relative to the A of the ATG codon/initiating methionine as the first nucleotide in the NCBI database (RefSeq accession NM_004448.1).

A role for the immunological synapse in lineage commitment of CD4 lymphocytes

Roberto A. Maldonado¹, Darrell J. Irvine², Robert Schreiber³ & Laurie H. Glimcher^{1,4}

¹Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts 02115-6017, USA

²Biological Engineering Division/Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

³Department of Pathology and Immunology, Center for Immunology, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA

⁴Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA

Activation of the naive T-helper lymphocyte (Thp) directs it down one of two major developmental pathways called Th1 and Th2. Signals transmitted by T cell, co-stimulatory and cytokine receptors control Thp lineage commitment but the mechanism by which these signals are integrated remains a mystery. The interferon- γ (IFNGR) and interleukin 4 (IL-4R) cytokine receptors, in particular, direct the earliest stages of T-helper commitment. Here we report that on engagement of the T-cell receptor (TCR) on Thp cells, rapid co-polarization of IFNGR with the TCR occurs within the developing immunological synapse. Thp cells from the intrinsically Th1-like C57BL/6 mouse strain have significantly more receptor co-polarization than Th2-prone BALB/c Thp cells. Remarkably, in the presence of IL-4, a cytokine required for Th2 differentiation, IFNGR co-polarization with TCR is prevented. This inhibition depends on Stat6, the transcription factor downstream of IL-4R that is required for Th2 differentiation. This cytokine receptor crossregulation provides an explanation for the effect of IL-4 in inhibiting Th1 differentiation. These observations suggest a scenario in which physical co-polarization of critical receptors directs the fate of the naive Thp, and offer a novel function for the immunological synapse in directing cell differentiation. They further suggest a new mechanism of membrane-bound signalling control by the physical disruption of large receptor-rich domains on signalling through a functionally antagonistic receptor.

The immunological synapse can be described as a superstructure initiated at the contact surface between the T cell and the antigen presenting cell (APC) that optimizes activation by congregating signalling molecules in specialized regions of the membrane of both cell types¹⁻⁴. Differences in calcium transport, assembly of adaptor protein/signalling molecules and receptor distribution into lipid raft structures have been described in mature Th1 and Th2 cells^{5,6} raising the possibility that fundamental differences in immunological synapse formation on the Thp might influence T helper lineage commitment by mechanisms as yet unknown.

TCR and IFNGR co-polarize at the immunological synapse on TCR engagement

Most evidence indicates that T helper differentiation *in vivo* is dictated by cytokines in concert with TCR signalling⁷⁻⁹. Autocrine or paracrine production of IFN γ and IL-4 are required for Th1 and Th2 commitment, respectively, in part through the induction by these two key cytokines of the T-bet (Th1) and GATA-3 (Th2) subset specific transcription factors. Mice lacking components of the IFN γ or IL-4 signalling pathways (including T-bet and GATA-3), lack Th1 or Th2 cells^{7,9}. We analysed the relative membrane topology of TCR and cytokine receptors, focusing on these two cytokine receptors (IFNGR and IL-4R) on the naive Thp, that are critical for the earliest stages of T-helper commitment. Many previous studies have been performed using polyclonally activated T cells and transformed B cells as APC. To more closely mirror physiological conditions as observed *in vivo*¹⁰, we studied the formation and evolution of the immunological synapse using Thp and mature splenic dendritic cells (DC). Two methodologies were pursued with similar outcomes. First, the cell suspension was sedimented by centrifugation to optimize interaction between the two cell types (almost 100% of T cells form conjugates with dendritic

cells), the reaction stopped by fixation with paraformaldehyde (PFA), Thp cell-DC couplets permeabilized, labelled and observed by confocal microscopy. Analysis of receptor distribution on T cells demonstrated that at 0 min the TCR β , IFNGR1 (IFNGR α subunit and IFNGR2, the β subunit) and IL-4R (data not shown) were randomly distributed on the cell surface with little or no receptor accumulation at the Thp cell-DC interface (Fig. 1a; and Supplementary Fig. 1). In contrast, if cells were incubated 30 min (and 10 min, not shown) before fixation, 60 to 80% of the 50 cells analysed per experiment ($n = 4$) displayed increased polarization of TCR β in the presence of ovalbumin (OVA)-loaded dendritic cells (Fig. 1a) but not with unpulsed dendritic cells (data not shown). Strikingly, IFNGR1 (and IFNGR2, Supplementary Fig. 1) was recruited to the Thp cell-DC interface and colocalized with TCR β . As reported, activation of the Thp leads to patching and internalization of the TCR. However, this is a continuous and not completely synchronized process. Given that polarization and subsequent internalization occur in the same geographical location, we chose 30 min, a time when the majority of Thp has achieved polarization (and some receptor internalization) to perform imaging studies. Further studies will be necessary to determine whether TCR and IFNGR receptors are internalized in the same vesicles or segregated after internalization.

Second, we performed four-dimensional (4D) imaging by mixing prelabelled transgenic DO11.10 Thp cells and OVA-loaded dendritic cells in order to visualize better the dynamics of DC-Thp cell interactions and receptor movement relative to the Ca²⁺ influx that follows T cell activation (assessed by FURA-2AM radiometric analysis). A representative experiment (out of four) of time-lapse microscopy in Fig. 1b-d (and Supplementary Movies 1 and 2) shows that at 0 s (no cell-cell contact) Ca²⁺ levels were low and both TCR β and IFNGR1 were uniformly distributed in the cell

membrane. Immediately after Ca^{2+} influx, both IFNGR1 and TCR β on the naive T cell progressively migrate to the contact interface with the dendritic cell (Fig. 1b, c).

Restricted co-polarization of IFNGR and IL-2R cytokine receptors with the TCR at the immunological synapse

We obtained similar results when Thp activation was induced by crosslinking of TCR at 37°C with a secondary antibody in the absence of dendritic cells. Figure 2a shows that crosslinking of TCR β induced its segregation and co-polarization with the IFNGR1 (and IFNGR2, not shown) but not IL-4R. In order to visualize better the co-polarization of receptors by a more quantitative method, values of their fluorescence intensity were measured (or 'linearized'). The left panels in Fig. 2c show that the distribution of TCR β and IFNGR1 molecules at 0 min is random and uniformly represented across the entire cell surface. At 30 min and 10 min (data not shown) after TCR crosslinking both TCR β and IFNGR1 molecules were polarized to one region at the periphery of the cell (Fig. 2a) as reflected by the one-peak shape of the almost completely overlapping curves (Fig. 2c, first row, right panel). In this system TCR-induced copatching of cytokine receptors was quite specific to the IFNGR as the IL-4R, IL-6R, IL-7R and IL-10R (Fig. 2b, c) did not colocalize with TCR and their distribution was similar to CD45, a molecule known to be excluded from the immunological synapse³. The single exception was the IL-2R alpha chain, previously noted to be present in membrane rafts¹¹, which displayed a similar (although less pronounced) co-polarization with the TCR after activation. This is interesting because IL-2, the earliest cytokine secreted by the Thp, is an overall growth factor for both T-helper subsets but is especially important for Th2 development¹². Because ligand-induced IFNGR polarization to caveolar membrane domains and

internalization occurs¹³, the patching observed could be a consequence of small quantities of IFN γ secreted by contaminating mature T-helper cells or by early activated Thp. This was not the case, however, as a comparison of TCR crosslinking on WT, IFN γ -deficient and Stat1-deficient Thp showed an identical pattern of distribution and motility of IFNGR1 and TCR (Fig. 2c, bottom row and Supplementary Fig. 2). Of note, neither the presence of IL-4 nor the absence of IFN γ (using IFN γ -deficient Thp as above) induced the co-polarization of TCR and IL-4R (Supplementary Fig. 3). We conclude that activation of the naive Thp results in a rapid and selective co-polarization of TCR and IFNGR and that this co-polarization is IFN γ -independent.

Concordance between the intrinsic Th1-'ness' of a Thp cell and the degree of IFNGR and TCR co-polarization

Genetic background controls the inherent tendency of Thp to differentiate along a Th1 or Th2 pathway, traits that have been mapped to specific chromosomal loci¹⁴. No firm explanation, however, has as yet emerged to explain such differences. We explored the possibility that Th1- or Th2-like strains of mice might display differences in their capacity to co-polarize IFNGR and TCR. In Fig. 3a, imaging analysis and linearization of the Thp surface (as in Fig. 2) shows that TCR crosslinking results in the extensive co-recruitment and polarization of IFNGR1 and TCR in Thp from the B6 (Th1-prone) strain. In contrast, in Thp from the BALB/c (Th2-prone) strain, the distribution of IFNGR is less polarized and its superimposition on the TCR only partial. In a more extensive analysis (Fig. 3b), the correlation coefficient ($\sigma_{x,y}$) was calculated for the normalized data (relative to the highest pixel peak) in order to determine whether the variations in the expression of IFNGR and TCR β at different sectors of the Thp membrane are

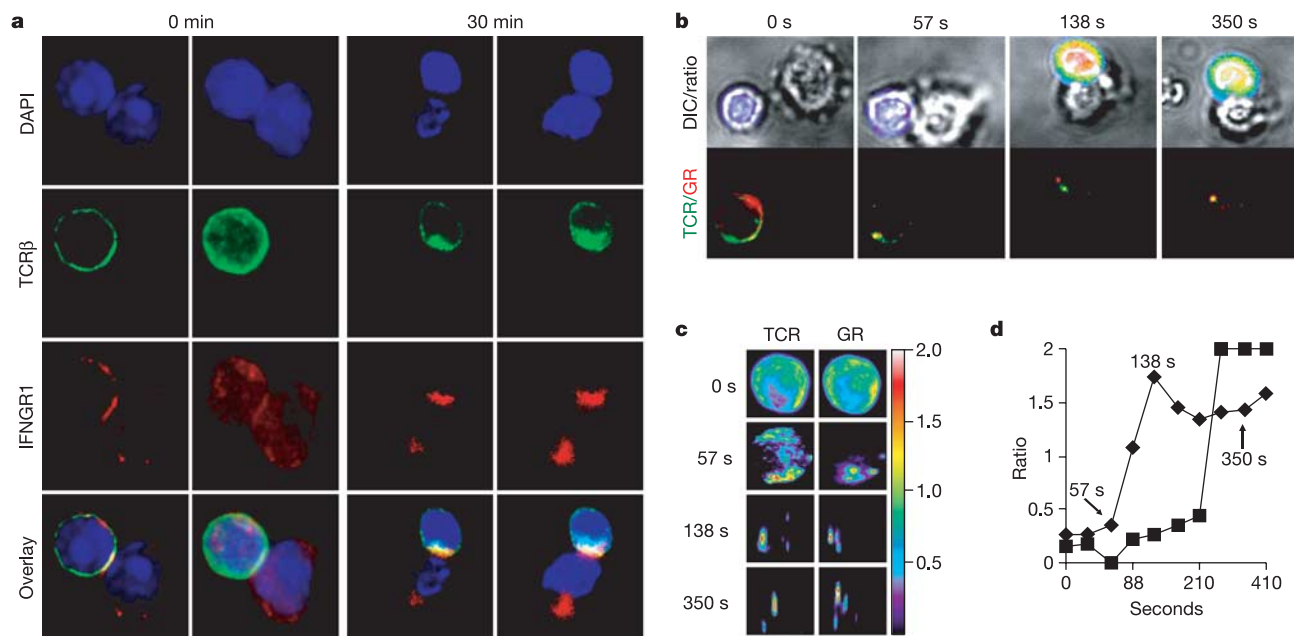


Figure 1 TCR β and IFNGR1 colocalize at the immunological synapse. **a**, CD62L^{high}CD4⁺ transgenic DO11.10 Thp and mature splenic CD11c⁺ OVA protein and peptide-pulsed dendritic cell sorted cells were spun together to maximize contact and the reaction stopped at sequential time points. Cells were permeabilized, stained using a combination of DAPI (blue) and directly-labelled monoclonal antibodies against TCR β (green) and IFNGR1 (red), IFNGR2 (far red, Supplementary Fig. 1) and observed by confocal microscopy. At each time point, the middle plane of the Z axis (left panels) and maximum projection of 25 Z planes (right panels) of the cell are shown. **b–d**, 4D observation of the dynamics of Thp activation and receptor motility in the presence of protein/peptide-loaded dendritic cells (Supplementary Movie 2). Thp were stained using the same combination of

membrane markers as above (TCR β in green, IFNGR1 in red). Additionally, calcium content measurements were performed by radiometric analysis of Fura-2AM fluorescence (see Methods). DIC images of Thp cell–DC interaction superimposed onto the false colour representation of Fura-2AM radiometric analysis (overlay, top panels) and the staining of TCR β and IFNGR1 (overlay, bottom panels) (**b**). A false colour reconstruction of anti-TCR β (left panels) and anti-IFNGR1 (right panels) fluorescence observed from the viewpoint of the dendritic cell looking towards the Thp (**c**). Plot of time-course evolution of the intracellular calcium concentration in the responding T cell (diamonds). As a control cells were cultured in the presence of 1 μM ionomycin to induce a strong calcium influx (squares) (**d**).

linked. The value of $\sigma_{x,y}$ increased up to fivefold from 0 min to 30 min after Thp stimulation, reflecting the active recruitment of both receptors. The chance of TCR/IFNGR co-polarization occurring at random in the B6 Thp is approximately 1/5000 (average $\sigma_{x,y}$ value of 0.9; $P = 0.0002$) whereas it is approximately 1/25 (average $\sigma_{x,y}$ value of 0.58; $P = 0.04$) in the BALB/c Thp, a 200-fold

difference. These results indicate a concordance between the tendency of a Thp to differentiate along the Th1 pathway and the degree of IFNGR and TCR co-polarization. This was not directly attributable to IL-4 itself as IL-4^{-/-} BALB/c Thp behaved similarly (Supplementary Fig. 4) suggesting that another intrinsic genetic mechanism was operative in the BALB/c strain. The less robust recruitment of

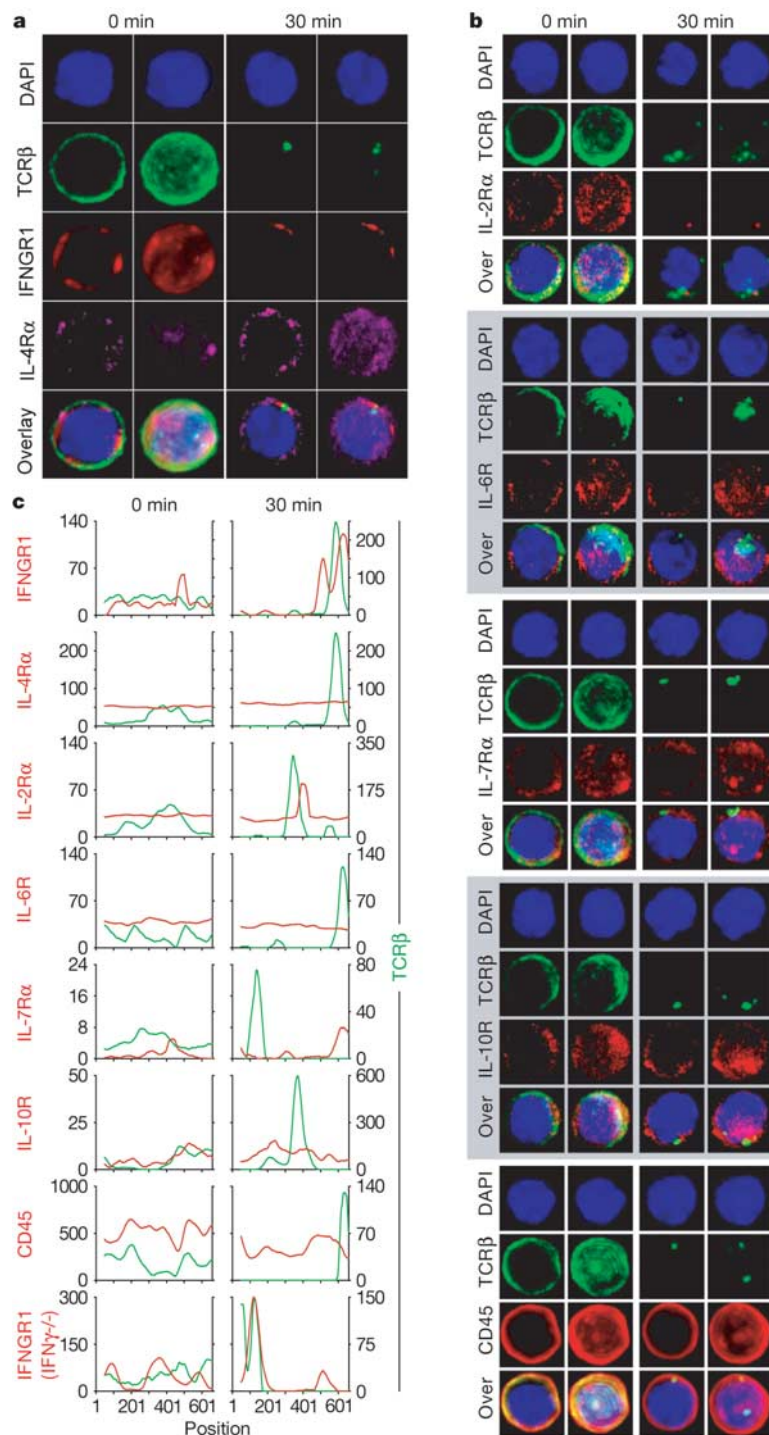


Figure 2 IFNGR co-polarization and partial IL-2R co-polarization with the TCR after T cell activation. Thp were activated by TCR crosslinking fixed and stained (as in Fig. 1a) using different combinations of directly labelled monoclonal antibodies as indicated. Each figure corresponds to one cell representative of 10 to 50 cells, from 3 different experiments. **a**, IFNGR but not IL-4R co-polarize with the TCR after Thp activation. The left panels represent the middle Z section of the cell whereas the right column depicts the maximum projection of all the (25) Z sections (see Fig. 1). **b**, Only IFNGR and IL-2Rα but not IL-6R,

IL-7R, IL-10R and CD45 co-polarize with the Thp after activation. Panels as in **a**. **c**, Linearization of the fluorescence on T cell membranes. The large number of points on the X axis are not relevant and correspond to the resolution of the pictures (600 to 700 points in a circumference of approximately 30 nm). The fluorescence of this region was plotted as number of pixels (Y axis) relative to their position along the region (X axis). The bottom plot depicts the same type of experiments using Thp from IFN γ -deficient mice.

IFNGR to the immunological synapse in the BALB/c Th2-like Thp led us to test directly whether differential recruitment might be important for Th1/Th2 differentiation.

IL-4 impedes the co-polarization of IFNGR with TCR in a Stat6-dependent manner

T helper differentiation is a rather unique system where a relatively neutral signal (TCR engagement) is accompanied by positive polarizing signals (cytokine receptor engagement) that are also mutually inhibitory⁸. Thus IL-4 inhibits the differentiation of Th1 cells from Thp and tends to be dominant over the Th2-inhibitory properties of the IFN γ Th1-enhancing cytokine. The mechanism by which such inhibition occurs is unknown. We tested the hypothesis that ligand-induced signalling through the IL-4R might affect the physical distribution of the opposing IFNGR. Thp were cultured under Th2 polarizing conditions and receptor localization was visualized. Remarkably, crosslinking of the TCR in the presence of IL-4 completely inhibited the migration and co-polarization of the IFNGR1 (Fig. 4a, and IFNGR2 not shown) with TCR. Population analysis (10 to 50 cells per experiment, $n = 3$) as above showed that the $\sigma_{x,y}$ values were significantly different when cultured in the presence of IL-4 (Fig. 4b). This was true when IL-4 was added at times ranging from 0 to 30 min after TCR crosslinking suggesting that IL-4 can actually reverse as well as prevent co-polarization. Other cytokines such as IL-2 and the Th1-antagonistic cytokine IL-10 did not interfere with co-polarization (Fig. 4c). Because IL-4 acts primarily through the Jak/Stat signalling pathway to accomplish Th2 lineage commitment, we tested the ability of IL-4 to inhibit TCR/IFNGR co-polarization in mice lacking the IL-4R signalling molecule, Stat6. Figure 4d shows that the ability of IL-4 to prevent TCR/IFNGR co-polarization is abolished in Stat6^{-/-} Thp cells. Occupancy of IL-4R with subsequent activation of Stat6 therefore translates into events that prevent the IFNGR from migrating into the immunological synapse with TCR. There is precedence for the ability of an inhibitory receptor to prevent the recruitment of an activating receptor into the immunological synapse. Engagement of the natural killer (NK) inhibitory receptors, killer cell immunoglobulin-like receptor (KIR) 2DL1 or CD94/NKG2, rapidly blocks tyrosine phosphorylation of the (CD244)

2B4 activating receptor by preventing its recruitment into lipid rafts¹⁵. However, it was not clear from those studies whether NK inhibitory receptors blocked recruitment because synapse formation was deficient, similar to the blockade of phosphorylation and Ca²⁺ flux by TGF β (ref. 16), or whether this was a specific process.

Discussion

The function of the T cell synapse remains controversial although there is general agreement that sustained signalling, which correlates with immunological synapse formation^{17,18}, leads to optimal T cell activation characterized by cycles of receptor signalling and degradation¹⁹. A recently espoused view that is consistent with most available data envisions the immunological synapse as a physical structure with the required complexity to guide T cell activation under diverse settings over time¹. Our data suggest that the IFNGR is a novel member of that architectural platform whose co-polarization with TCR may guide lineage commitment. One function of such polarization might include the concentration of cytokine receptors at the actual sites where cytokines are produced. This hypothesis fits nicely with previous contentions that cytokines act locally rather than at a distance²⁰ and assumes directional cytokine secretion focused towards the presentation pole of the cell, a phenomenon already described for CTLs (cytotoxic T lymphocytes) and NK cells²¹. On target recognition, these cells polarize the secretion of cytolytic factors with a 'kiss of death'²¹. Hence, for naive Thp the immunological synapse could provide a platform for 'cytokine presentation' of key activating and polarizing factors such as IL-2, IFN γ , IL-12, IL-23 or IL-27 from dendritic cells. Such presentation might be reciprocal to include cytokines presented from the T cell to the dendritic cell. Indeed, activated mature TCR transgenic T cells co-polarize TCR and IL-2 secretory vesicles *in situ* on injection of specific peptide suggesting secretion of IL-2 towards the immunological synapse²². Further, this autoactivating loop might also be initiated by dendritic cells, which secrete IL-2 shortly after maturation, another reason for the T cell to bring IL-2R to that interface²³. Thus, polarized IL-2R (as shown above) and IL-2 production will not only maximize the efficiency of IL-2 autoactivation but will also ensure optimum activation of APC presentation capacity.

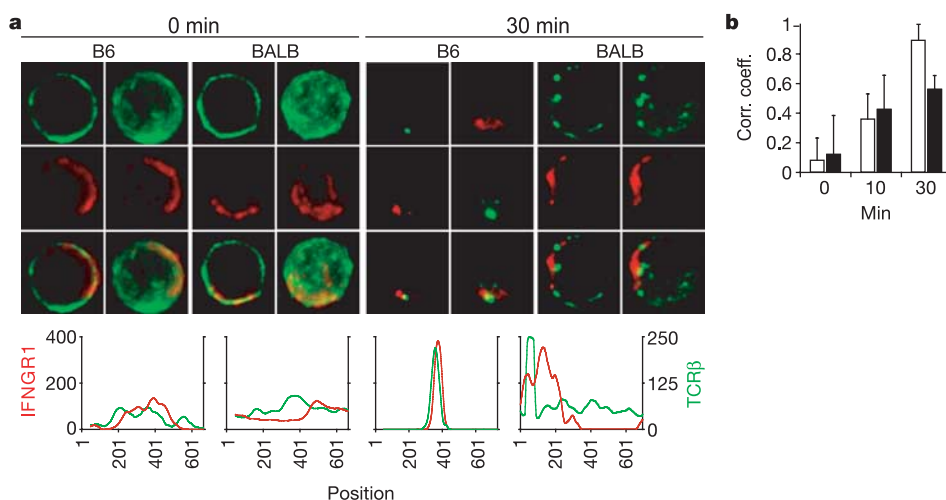


Figure 3 B6 Thp display increased co-polarization of TCR and IFNGR compared to BALB/c Thp. BALB/c and B6 Thp were activated by TCR crosslinking fixed and stained using a combination of directly labelled monoclonal antibodies (IFNGR1 in red, TCR β in green). **a**, Median sections (left panels) and maximum projections (right panels) of representative cells and the respective plots of their linearized surface (below). **b**, Correlation coefficient between IFNGR and TCR (see Methods). Shown is the average ($\pm$ s.d.) of the time lapse evolution of the correlation coefficients between the number of pixels of TCR and IFNGR

at the surface of Thp (10 to 50 per experiment; $n = 4$) from B6 mice (open bars) or BALB/c mice (filled bars). At 30 min the average of values of $\sigma_{x,y}$ are 0.58 and 0.9, respectively a 1.55-fold difference. Of note, an average of 2.5-fold decrease in the standard deviation of $\sigma_{x,y}$ values was observed from 0 min to 30 min reflecting the stabilization of TCR β /IFNGR co-polarization. Levels of IFNGR on B6 and BALB/c Thp were equivalent (not shown).

It is tempting to hypothesize that TCR-induced co-polarization leads to IFNGR signalling to preferentially drive Th1 differentiation. The presence of the IL-2R and IFNGR in the immunological synapse under 'neutral' conditions (the absence of IFN γ and IL-4) is consistent with previous reports that TCR signalling can induce Stat5 tyrosine²⁴ and Stat1 serine phosphorylation²⁵ and results in

transient downregulation of IL-4R signalling²⁶. However, the majority of these experiments involved strong TCR signalling and much evidence emphasizes the importance of cytokine signalling under TCR 'neutral' stimuli^{8,9}. In the Thp cell, signalling through the IFNGR and the IL-27R and Stat1, rather than the IL-12R (which is only expressed on activated T helper but not naive Thp cells) is

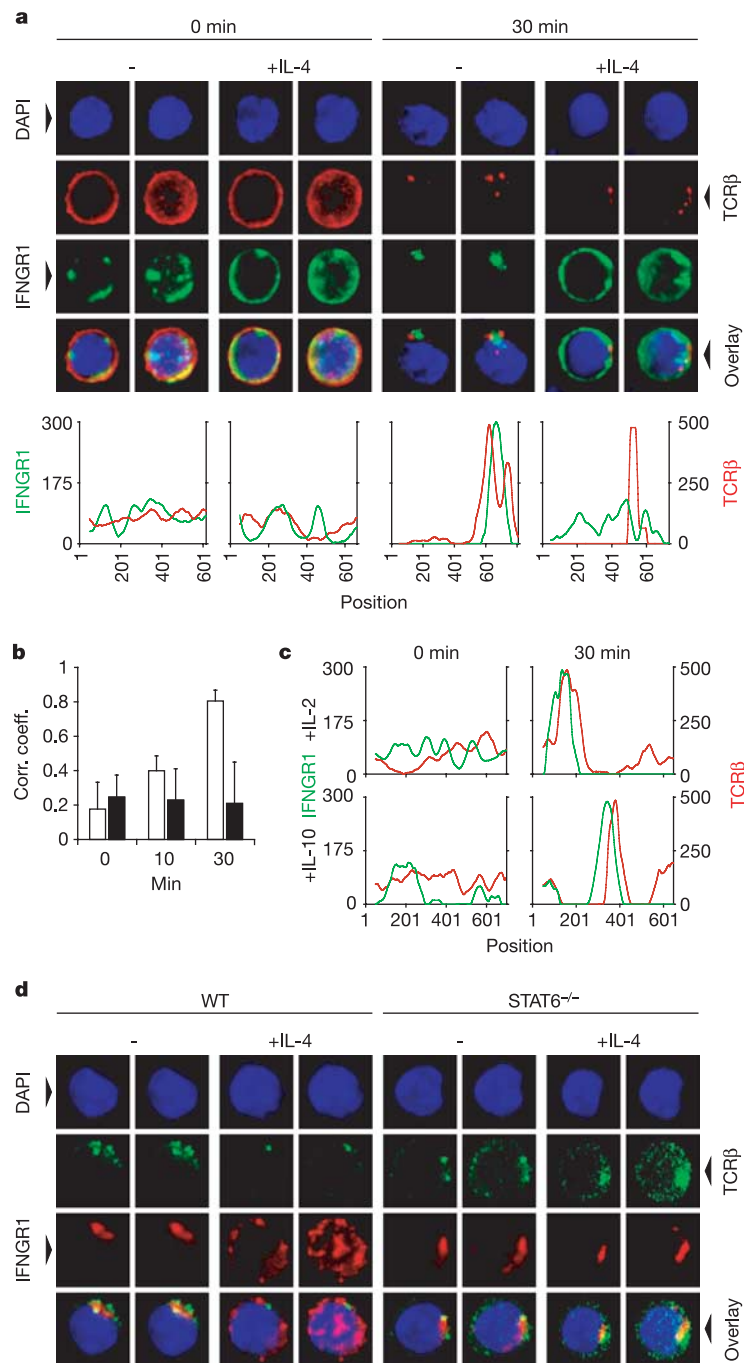


Figure 4 IL-4 impedes the co-polarization of IFNGR with TCR in a Stat6-dependent manner. **a**, Cells were cultured for various times with or without 20 ng ml⁻¹ of rIL-4 during TCR crosslinking, fixed at the indicated time and stained (as in Fig. 1a) using a combination of directly labelled monoclonal antibodies as indicated. Representative median single cell sections (left panels), maximum projections (right panels) and the corresponding cell surface linearizations plots (below) are shown (TCR in red, IFNGR1 in green). **b**, Correlation coefficients between the values of TCR and IFNGR (10 to 50 cells per experiment, $n = 5$). As in Fig. 3, the plot represents the average of individual correlation coefficients between the TCR and IFNGR ($\pm$ s.d.) for B6 Thp after TCR

crosslinking in the presence (filled bars) or absence (open bars) of IL-4. The average of $\sigma_{x,y}$ values from 0 min to 30 min varied on average from 0.18 to 0.8 when cells were untreated and from 0.25 to 0.21 when cultured in the presence of IL-4. **c**, Surface scan plots of Thp cultured in the presence of IL-2 and IL-10, for 0 min or 30 min after TCR crosslinking under the same conditions as previously described. **d**, Thp from wild type (WT) or Stat6-deficient animals were compared. Depicted are representative median single cell sections (left panels) and maximum projections (right panels) of 10 to 50 cells per condition (per experiment, $n = 3$).

probably the primary Th1-polarizing stimulus⁷. Indeed, we did not observe TCR/IFNGR co-polarization on mature T helper cells (data not shown). Our observations suggest a scenario where IFNGR co-recruitment provides the early link between TCR and cytokine signalling. Strong TCR signalling leads to accentuated IFNGR co-polarization and the assembly of a Th1 signalosome which is further stabilized by the subsequent secretion of IFN γ (ref. 27) unless an inhibitory signal (IL-4 induced Stat6 activation) is delivered and co-polarization prevented. Parallel engagement of TCR and IL-4R might then lead to assembly of a Th2 signalosome⁶. Such a scenario fits with the observation that activated Thp produce IL-4 promptly and can become Th2 cells in the absence of IL-12 or IFN γ (ref. 28). The quantity of IL-4 in the local milieu may ultimately determine Thp cell fate thus implying that the Th1 pathway may be the default response of the Thp. These speculations will need to be tested experimentally but our observations provide intriguing evidence for a new function of the immunological synapse in controlling cell fate decisions. It may be that similar processes occur for other receptors such as chemokine receptors whose residence in membrane microdomains with CD4 influences HIV binding²⁹ and in non-T cells where cytokine antagonism exists. Further, receptor co-polarization in cells of other lineages such as neurons, adipocytes, endocrine tissues and muscle—where lipid rafts have been described to control effector function—may also control cell fate³⁰. □

Methods

Cell preparations and Thp activation

Naive Thp cells were isolated from the lymph nodes of 5–6-week-old Rag2^{-/-} DO11.10 TCR transgenic mice by MACS (magnetic activated cell sorting) negative selection of CD25^{high}CD11c⁺CD8 α ⁺CD11b⁺CD45RB⁺DX5⁺CLII⁺ cells. Dendritic cells were purified by digesting the spleens of BALB/c mice with collagenase/dispase followed by MACS positive selection. CD11c⁺ sorted cells were cultured for 10 h to induce spontaneous maturation in the presence of 100 μ g ml⁻¹ of OVA protein and 1 μ M of OVA CLII-restricted peptide. The cell suspension was sedimented by centrifugation to optimize interaction between the two cell types (almost 100% of T cells form conjugates with dendritic cells), the reaction stopped by fixation with paraformaldehyde (PFA), Thp cell–DC couplets permeabilized, labelled and observed by confocal microscopy. In other experiments, crosslinking of the TCR was induced by adding a secondary goat anti-FITC Alexa 488-coupled antibody at 37 °C and the reaction was stopped by fixation of the cells at various times. Fixed cells were permeabilized and stained for different markers. 4D observations were performed by staining the cells with monoclonal antibody followed by culture at 37 °C in 5% CO₂ in separate droplets of complete media or three-dimensional (3D) collagen matrices (1 mg ml⁻¹). After 30 min rest (or collagen polymerization), a bridge between the droplets was created allowing Thp cell–DC cell contact. Markers used were as follows: DAPI, a DNA marker; Alexa-594-labelled (A594)- or A488-hamster anti-mouse IFN γ R α (2E2); A488- or A594-hamster anti-mouse TCR β (H57); A647-hamster anti-mouse IFN γ R β (MOB-47); A647-rat anti-mouse IL-4R (M1); (phycoerythrin (PE)-labelled) PE-rat anti-mouse IL-2R α (PC61); PE-rat anti-mouse IL-6R α (D7715A7); PE-rat anti-mouse IL-7R α (SB/14); PE-rat anti-mouse IL-10R (1B1.3a); PE-rat anti-mouse CD45RB (16A). Values of Fura-2AM radiometric analysis were obtained as Ratio = Fura emissions at 340 nm/380 nm.

Linearization analysis and calculation of correlation coefficient

Linearization analysis allows us to compare the location of the molecules observed in one plane of the cell versus the whole cell. Scans of the cell surface were made by drawing ring-shaped regions at the mid focal plane of every cell that included the membrane and the subjacent cytoplasm. The two arrays of fluorescence intensities per cell (TCR and IFNGR number of pixels) were normalized relative to the highest peak (Y axis) and their correlation coefficient (a statistic that measures the degree to which two variables are related, <http://noppa5.pc.helsinki.fi/koe/corr/cor7.html>, <http://mathworld.wolfram.com/CorrelationCoefficient.html>, $\sigma_{x,y}$) was calculated as $\rho_{x,y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$. This value is a version of the covariance expressed as $-1 \geq \sigma_{x,y} \geq +1$ and represents the relationship between two arrays of numbers where in 100 datapoints, a value of the correlation coefficient greater than 0.197 has less than a 5% probability of arising by chance.

Received 14 June; accepted 6 August 2004; doi:10.1038/nature02916.
Published online 22 September 2004.

- Huppa, J. B. & Davis, M. M. T-cell-antigen recognition and the immunological synapse. *Nature Rev. Immunol.* **3**, 973–983 (2003).
- Grakoui, A. *et al.* The immunological synapse: a molecular machine controlling T cell activation. *Science* **285**, 221–227 (1999).
- Bromley, S. K. *et al.* The immunological synapse. *Annu. Rev. Immunol.* **19**, 375–396 (2001).
- Viola, A., Schroeder, S., Sakakibara, Y. & Lanzavecchia, A. T lymphocyte costimulation mediated by reorganization of membrane microdomains. *Science* **283**, 680–682 (1999).
- Balamuth, F., Leitenberg, D., Unteraehrer, J., Mellman, I. & Bottomly, K. Distinct patterns of membrane microdomain partitioning in Th1 and Th2 cells. *Immunity* **15**, 729–738 (2001).
- Madrenas, J. A SLAT in the Th2 signalosome. *Immunity* **18**, 459–461 (2003).
- Szabo, S. J., Sullivan, B. M., Peng, S. L. & Glimcher, L. H. Molecular mechanisms regulating Th1 immune responses. *Annu. Rev. Immunol.* **21**, 713–758 (2003).
- O'Garra, A. Cytokines induce the development of functionally heterogeneous T helper cell subsets. *Immunity* **8**, 275–283 (1998).
- Murphy, K. M. & Reiner, S. L. The lineage decisions of helper T cells. *Nature Rev. Immunol.* **2**, 933–944 (2002).
- Miller, M. J., Wei, S. H., Parker, I. & Cahalan, M. D. Two-photon imaging of lymphocyte motility and antigen response in intact lymph node. *Science* **296**, 1869–1873 (2002).
- Vereb, G. *et al.* Cholesterol-dependent clustering of IL-2R α and its colocalization with HLA and CD48 on T lymphoma cells suggest their functional association with lipid rafts. *Proc. Natl Acad. Sci. USA* **97**, 6013–6018 (2000).
- Zhu, J., Cote-Sierra, J., Guo, L. & Paul, W. E. Stat5 activation plays a critical role in Th2 differentiation. *Immunity* **19**, 739–748 (2003).
- Takaoka, A. *et al.* Cross talk between interferon-gamma and -alpha/beta signaling components in caveolar membrane domains. *Science* **288**, 2357–2360 (2000).
- Fowell, D. J. & Locksley, R. M. *Leishmania major* infection of inbred mice: unmasking genetic determinants of infectious diseases. *Bioessays* **21**, 510–518 (1999).
- Watzl, C. & Long, E. O. Natural killer cell inhibitory receptors block actin cytoskeleton-dependent recruitment of 2B4 (CD244) to lipid rafts. *J. Exp. Med.* **197**, 77–85 (2003).
- Chen, C. H. *et al.* Transforming growth factor beta blocks Tec kinase phosphorylation, Ca²⁺ influx, and NFATc translocation causing inhibition of T cell differentiation. *J. Exp. Med.* **197**, 1689–1699 (2003).
- Lee, K. H. *et al.* T cell receptor signaling precedes immunological synapse formation. *Science* **295**, 1539–1542 (2002).
- Huppa, J. B., Gleimer, M., Sumen, C. & Davis, M. M. Continuous T cell receptor signaling required for synapse maintenance and full effector potential. *Nature Immunol.* **4**, 749–755 (2003).
- Lee, K. H. *et al.* The immunological synapse balances T cell receptor signaling and degradation. *Science* **302**, 1218–1222 (2003).
- Poo, W. J., Conrad, L. & Janeway, C. A. Receptor-directed focusing of lymphokine release by helper T cells. *Nature* **332**, 378–380 (1988).
- Trambas, C. M. & Griffiths, G. M. Delivering the kiss of death. *Nature Immunol.* **4**, 399–403 (2003).
- Reichert, P., Reinhardt, R. L., Ingulli, E. & Jenkins, M. K. Cutting edge: *in vivo* identification of TCR redistribution and polarized IL-2 production by naive CD4 T cells. *J. Immunol.* **166**, 4278–4281 (2001).
- Granucci, F. *et al.* Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nature Immunol.* **2**, 882–888 (2001).
- Welte, T. *et al.* STAT5 interaction with the T cell receptor complex and stimulation of T cell proliferation. *Science* **283**, 222–225 (1999).
- Gamero, A. M. & Larner, A. C. Signaling via the T cell antigen receptor induces phosphorylation of Stat1 on serine 727. *J. Biol. Chem.* **275**, 16574–16578 (2000).
- Zhu, J. *et al.* Transient inhibition of interleukin 4 signaling by T cell receptor ligation. *J. Exp. Med.* **192**, 1125–1134 (2000).
- Zhang, Y. *et al.* Interferon gamma stabilizes the T helper cell type 1 phenotype. *J. Exp. Med.* **194**, 165–172 (2001).
- Laumois, P. *et al.* IL-4 rapidly produced by V β 4 V α 8 CD4⁺ T cells instructs Th2 development and susceptibility to *Leishmania major* in BALB/c mice. *Immunity* **6**, 541–549 (1997).
- Manes, S. *et al.* Membrane raft microdomains in chemokine receptor function. *Semin. Immunol.* **13**, 147–157 (2001).
- Tsui-Pierchala, B. A., Encinas, M., Milbrandt, J. & Johnson, E. M. Jr Lipid rafts in neuronal signaling and function. *Trends Neurosci.* **25**, 412–417 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Shaw, G. Petsko, M. Grusby and G. Lord for thoughtful review of the manuscript, and G. Petsko for advice on the statistical analysis. This work was supported by National Institutes of Health grants. R.M. is a recipient of the Kelli and Gerald Ford Irvington Institute Postdoctoral Fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.H.G. (lglimche@hsph.harvard.edu).

A supernova origin for dust in a high-redshift quasar

R. Maiolino¹, R. Schneider^{1,3}, E. Oliva^{1,4}, S. Bianchi², A. Ferrara⁵, F. Mannucci², M. Pedani⁴ & M. Roca Sogorb⁶

¹INAF - Osservatorio Astrofisico di Arcetri, ²CNR-IRA, Sezione di Firenze, Largo Enrico Fermi 5, 50125 Firenze, Italy

³"Enrico Fermi" Center, Via Panisperna 89/A, 00184 Roma, Italy

⁴Telescopio Nazionale Galileo, C. Alvarez de Abreu, 70, 38700 Santa Cruz de la Palma, Spain

⁵SISSA/International School for Advanced Studies, Via Beirut 4, 34100 Trieste, Italy

⁶Astrofisico Fco. Sánchez, Universidad de La Laguna, 38206 La Laguna, Tenerife, Spain

Interstellar dust plays a crucial role in the evolution of the Universe by assisting the formation of molecules¹, by triggering the formation of the first low-mass stars², and by absorbing stellar ultraviolet–optical light and subsequently re-emitting it at infrared/millimetre wavelengths. Dust is thought to be produced predominantly in the envelopes of evolved (age >1 Gyr), low-mass stars³. This picture has, however, recently been brought into question by the discovery of large masses of dust in the host galaxies of quasars^{4,5} at redshift $z > 6$, when the age of the Universe was less than 1 Gyr. Theoretical studies^{6–8}, corroborated by observations of nearby supernova remnants^{9–11}, have suggested that supernovae provide a fast and efficient dust formation environment in the early Universe. Here we report infrared observations of a quasar at redshift 6.2, which are used to obtain directly its dust extinction curve. We then show that such a curve is in excellent agreement with supernova dust models. This result demonstrates a supernova origin for dust in this high-redshift quasar, from which we infer that most of the dust at high redshifts probably has the same origin.

Powerful quasars offer an ideal tool for detailed studies of dust in their host galaxies, out to very high redshifts. Dust extinction is inferred through reddening of the quasar ultraviolet/optical continuum emission. The extinction curve ($A_\lambda = 1.086 \tau_\lambda$, where τ_λ is the wavelength-dependent optical depth) of dust associated with low redshift, mildly obscured quasars has been typically found to be consistent with that of the Small Magellanic Cloud (SMC)^{12–14}, while for heavily absorbed quasars there are indications that the extinction curve may be different^{15,16}.

An important class of quasars are the broad absorption line (BAL) quasars, whose ultraviolet spectrum is characterized by blueshifted, deep and broad absorption features associated with highly ionized atomic species, which trace powerful outflows of dense gas along our line of sight. Low-ionization BAL (LoBAL) quasars are characterized by additional low-ionization absorption lines, probably associated with higher column densities of gas. LoBAL quasars at $z < 4$ are always significantly reddened by dust (associated with the outflowing gas), and therefore they are ideal laboratories in which to investigate the properties of dust.

By means of low-resolution near-infrared spectroscopic observations¹⁷ we have recently identified a few BAL quasars at $z \approx 5$ –6. The most distant among them is the LoBAL quasar SDSSJ104845.05 + 463718.3 (hereafter SDSS1048 + 46). At a redshift $z = 6.193$ (based on a new medium-resolution spectrum around MgII λ 2798; unpublished work), this is the only LoBAL at $z > 5$ currently known, and it offers a unique chance of investigating the dust extinction curve at $z \approx 6$. This quasar was re-observed at higher spectral resolution with the goal of de-blending the deep troughs characterizing its spectrum and of estimating the reddening of the continuum. Observations were obtained in

February 2004 with the Near Infrared Camera Spectrometer (NICS¹⁸) at the Telescopio Nazionale Galileo in the Canary Islands. SDSS1048 + 46 was observed for a total of three hours, with a slit of 1.5" oriented along the parallactic angle. We used a grism covering the spectral region 0.8–1.45 μm .

Figure 1a shows the observed spectrum (solid line) smoothed to a lower resolution for sake of clarity, and combined with a previous low-resolution spectrum¹⁷ of the full wavelength range 0.8–2.4 μm . At the quasar's redshift our data cover the wavelength range $1,200 < \lambda_{\text{rest}} < 3,300 \text{ \AA}$ in the rest frame of the emitted radiation. The observed spectrum is much bluer than LoBAL quasars observed at $z < 4$, whose average spectrum is shown with the dashed line¹³. In particular, SDSS1048 + 46 is bluer than any known LoBAL quasar¹³ at $z < 4$. Because the red shape of LoBAL quasars at $z < 4$ is ascribed to dust reddening, the much bluer slope of SDSS1048 + 46 already indicates a large variation of dust extinction and reddening from $z < 4$ to $z \approx 6$.

The continuum at $\lambda_{\text{rest}} > 1,700 \text{ \AA}$ can be fitted with a non-BAL (unreddened) quasar template with a slope $\alpha = -2.1$ ($F_\lambda \propto \lambda^\alpha$), which is shown with a dotted line in Fig. 1. Such a blue slope is consistent with the intrinsic slope of -2.01 inferred by ref. 13 for BAL quasars. The observed spectrum deviates significantly from the non-BAL unreddened template only at $\lambda_{\text{rest}} < 1,700 \text{ \AA}$. While such a deviation was initially ascribed to a severe blend of the CIV and SiIV troughs in the low-resolution spectrum, the new medium-resolution spectrum (whose unsmoothed and enlarged version is shown in Fig. 1b) clearly shows that this is not the case. The continuum outside the troughs is indeed redder than expected by the extrapolation of the longer wavelength spectrum through the unreddened quasar template (dotted line). If such reddening at short wavelengths is due to dust, then the extinction curve must be quite unusual: relatively flat at $\lambda > 1,700 \text{ \AA}$ and steeply rising at shorter wavelengths. We can quantitatively derive the extinction curve by using the equation $A_\lambda = -2.5 \log(F_{\text{obs}}/F_{\text{intr}})$, where F_{obs} is the observed spectrum and F_{intr} is the intrinsic spectrum. We avoided spectral regions contaminated by emission/absorption features by interpolating the continuum from the nearby regions. The blend of Fe II lines¹⁹ makes the description of the continuum in the spectral region between 2,300 $\text{\AA}$ and 3,050 $\text{\AA}$ more uncertain; however, the latter spectral region is not critical within the context of this paper, because the most important aspect is that the extinction curve must be rather flat in the region between 1,700 $\text{\AA}$ and 3,300 $\text{\AA}$, and increase rapidly at $\lambda_{\text{rest}} < 1,700 \text{ \AA}$. For the intrinsic spectrum we used the non-BAL template obtained by the SLOAN survey¹³ (this choice is appropriate because non-BAL quasars at $z \approx 6$ have spectra similar to those at $z < 4$)¹⁷, with slopes ranging from $\alpha = -2.1$ (the reddest slope compatible with our observed spectrum at $\lambda_{\text{rest}} > 1,700 \text{ \AA}$) to $\alpha = -2.5$ (only 1% of quasars have slopes bluer than this value¹³). The resulting extinction curve is shown in Fig. 2 (thick solid line and shaded region). As expected, the extinction curve inferred for this quasar at $z = 6.2$ is quite different with respect to the SMC curve which applies to quasars at $z < 4$. The dot-dashed line in Fig. 1b indicates the non-BAL template absorbed with the extinction curve inferred by us in Fig. 2, which nicely matches the observed spectrum (except for the emission and absorption lines, whose intensity and shape depend on the physics of the ionized gas). The extinction $A_{3,000 \text{ \AA}}$ required to match the observed spectrum is in the range 0.4–0.8 mag.

The extinction curve inferred from the most distant LoBAL can also reproduce the shape of the second-most-distant BAL SDSS1044-01, a high-ionization BAL (HiBAL)¹⁷ at $z = 5.78$. The latter quasar presents much lower reddening than SDSS1048 + 46 and therefore cannot be used to provide tight constraints on the extinction curve. Nonetheless, similarly to SDSS1048 + 46, it is characterized by a rather blue continuum at $\lambda_{\text{rest}} > 1,700 \text{ \AA}$ and a reddening at $\lambda_{\text{rest}} < 1,700 \text{ \AA}$, which can be nicely fitted by using the same extinction curve derived above and shown in Fig. 2.

In order to interpret the observations, we have computed the supernova dust extinction curve by using the model proposed by ref. 6, which describes dust formation in the ejecta of Type-II supernovae as a function of the progenitor mass and metallicity. In spite of the uncertainties, the model has been successfully applied

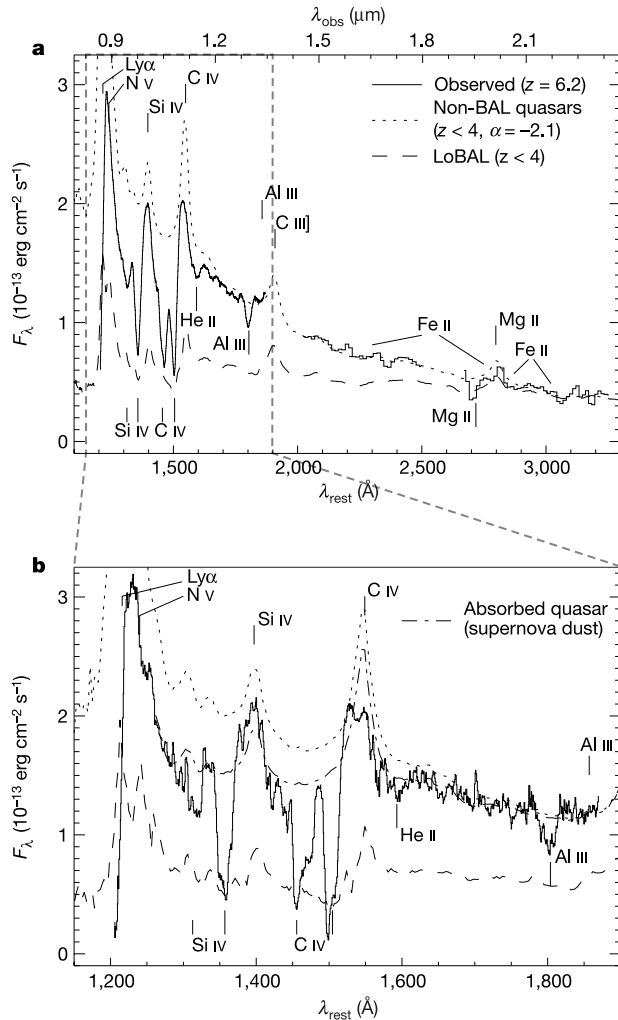


Figure 1 Spectrum of the quasar SDSS1048 + 46 at redshift 6.2, compared with quasar templates at lower redshift. **a**, The solid line shows the low-resolution near-infrared spectrum of the LoBAL quasar SDSS1048 + 46. It is the composition of a previous low-resolution $\lambda/\Delta\lambda = 75$ spectrum¹⁷ (shown only for $\lambda > 1.4 \mu\text{m}$), and the new medium-resolution $\lambda/\Delta\lambda = 350$ spectrum smoothed to a resolution of 75 for sake of clarity. Cross-calibration between the two spectra is ensured by matching the common parts. Upper labels identify the emission lines while lower labels identify the blueshifted absorption features. The presence of strong absorption from the high-ionization species C IV $\lambda = 1,549$ and Si IV $\lambda = 1,397$, along with absorption by the low-ionization species Al III $\lambda = 1,857$ and Mg II $\lambda = 2,798$ classifies this as a LoBAL quasar. The missing parts of the spectrum are due to the regions of bad atmospheric transmission. The dotted line is the average spectrum of (unreddened) non-BAL quasars at $z < 4$ with a slope $\alpha = -2.1$. The dashed line is the average spectrum of LoBAL quasars at $z < 4$. **b**, Same as above, but where the solid line shows only the (unsmoothed) medium resolution spectrum ($\lambda/\Delta\lambda = 350$) of SDSS1048 + 46. It is important to note the continuum level outside the troughs at $\lambda < 1,700 \text{ Å}$, which is well below the extrapolation of the continuum at longer wavelengths in the case of no dust reddening (dotted line). Both low- and high-resolution spectra allow the identification of the spectral regions for the continuum fitting (required to derive the extinction curve), and more specifically: 1,275–1,290 Å, 1,325–1,304 Å, 1,417–1,434 Å, 1,482–1,489 Å, 1,577–1,590 Å, 1,567–1,677 Å, 1,708–1,782 Å, 1,815–1,845 Å, 2,015–2,270 Å, 3,050–3,255 Å. The dot-dashed line in **b** shows the effect of absorbing the non-BAL template with the extinction curve inferred by us and shown in Fig. 2 (thick solid line).

to interpret the observed properties of SN 1987A (ref. 6) and of the young, metal-poor dwarf galaxy SBS 0335-052 (ref. 20). A grid of supernova models has been considered, with different initial stellar progenitor masses ($12M_{\odot} \leq M \leq 40M_{\odot}$) and metallicities ($0 \leq Z/Z_{\odot} \leq 1$)²¹. The resulting supernova dust grains are made of silicates (Mg_2SiO_4 and MgSiO_3), amorphous carbon, magnetite (Fe_3O_4) and corundum (Al_2O_3). The typical grain sizes range between 5 Å and 0.1 μm , with smaller grains being predominantly composed of silicates and magnetite (sizes about 5–15 Å) and larger ones by amorphous carbon (sizes about 100–200 Å). The predicted grain properties for different Type-II supernovae have then been averaged over the stellar Initial Mass Function (IMF). We adopted a Salpeter IMF, but even higher-mass weighted Larson-like IMFs²² do not change the results significantly ($\lesssim 2\%$ in the extinction curve).

The supernova dust extinction properties have been derived using the standard Mie theory for spherical grains. Grains made of amorphous carbon are the main contributors to extinction. We used the optical properties for amorphous carbon produced in an inert atmosphere (ACAR²³), which is appropriate for the supernova ejecta environment. Other important contributors to extinction were found to be Mg_2SiO_4 and Fe_3O_4 grains^{24,25}.

Figure 2 shows the resulting extinction curves, which are in much better agreement with the observations than the usual SMC curve. The extinction curves were obtained for various initial stellar metallicities. The best agreement is found for the $Z = 10^{-2}Z_{\odot}$ and $10^{-4}Z_{\odot}$ models, although the other models do not differ strongly and are still within the observational uncertainties. We also show the case of dust formed in the ejecta of a single $25M_{\odot}$, $Z = 10^{-4}Z_{\odot}$ Type-II supernova; the agreement for this case is

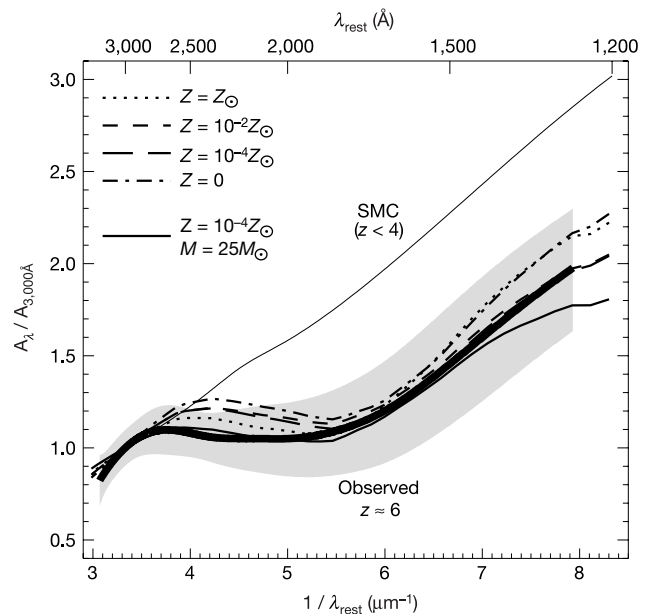


Figure 2 Extinction curve observed in the quasar SDSS1048 + 46 at $z = 6.2$ compared with the extinction curve observed in quasars at $z < 4$ and with the extinction curve expected from supernova dust. The thick solid line shows the extinction curve inferred for SDSS1048 + 46. The shaded area shows the associated uncertainty, which includes the range of intrinsic spectral slopes (this uncertainty dominates at short wavelengths, $\lambda < 1,700 \text{ Å}$), the uncertainty on the continuum interpolation (in particular around the Fe II hump between 2,300 Å and 3,050 Å), and the noise in the spectrum. The thin solid line shows the SMC extinction curve, which applies to quasars at $z < 4$ (including BAL quasars). The other lines indicate theoretical predictions for dust produced by supernovae: the dot-dashed, long-dashed, short-dashed and dotted lines represent the extinction curves produced by supernovae obtained assuming that stars form according to a Salpeter IMF and initial metallicities 0, 10^{-4} , 10^{-2} , and $1Z_{\odot}$, respectively. We also show the case of dust formed in the ejecta of a single $25M_{\odot}$, $Z = 10^{-4}Z_{\odot}$ Type-II supernova (medium solid line).

excellent at most wavelengths. The plateau between 1,700 Å and 3,000 Å is due to a local minimum between two broad absorption features caused by amorphous carbon grains, with the added flat contribution from Fe₃O₄ grains. Silicate grains contribute to the rise at shorter wavelengths.

We conclude that our analysis, combined with observations, provides the first direct evidence for dust produced in supernova ejecta, rather than in evolved stars, in an object at $z > 6$. In particular, we find that dust purely produced by Type-II supernovae can explain the observed extinction curve very well. By implication, we conclude that much of the dust seen at high redshifts probably has the same origin. The properties of high-redshift dust are therefore noticeably different from those found at later cosmic times: grains are typically smaller, owing to their different formation history and to the short time available to subsequently accrete heavy atoms and coagulate with other grains. □

Received 22 April; accepted 6 August 2004; doi:10.1038/nature02930.

- Hirashita, H. & Ferrara, A. Effects of dust grains on early galaxy evolution. *Mon. Not. R. Astron. Soc.* **337**, 921–937 (2002).
- Schneider, R., Ferrara, A., Salvaterra, R., Omukai, K. & Bromm, V. Low-mass relics of early star formation. *Nature* **422**, 869–871 (2003).
- Whittet, D. C. B. *Dust in the Galactic Environment* (Series in Astronomy & Astrophysics, Institute of Physics (IOP) Publishing, Bristol, 2003).
- Bertoldi, F. *et al.* Dust emission from the most distant quasars. *Astron. Astrophys.* **406**, L55–L58 (2003).
- Priddy, R. S., Isaak, K. G., McMahon, R. G., Robson, E. I. & Pearson, C. P. Quasars as probes of the submillimetre cosmos at $z > 5$. I. Preliminary SCUBA photometry. *Mon. Not. R. Astron. Soc.* **344**, L74–L78 (2003).
- Todini, P. & Ferrara, A. Dust formation in primordial Type II supernovae. *Mon. Not. R. Astron. Soc.* **325**, 726–736 (2001).
- Nozawa, T., Kozasa, T., Umeda, H., Maeda, K. & Nomoto, K. Dust in the early Universe: dust formation in the ejecta of population III supernovae. *Astrophys. J.* **598**, 785–803 (2003).
- Schneider, R., Ferrara, A. & Salvaterra, R. Dust formation in very massive primordial supernovae. *Mon. Not. R. Astron. Soc.* **351**, 1379–1386 (2004).
- Moseley, S. H., Dwek, E., Glaccum, W., Graham, J. R. & Lowenstein, R. F. Far-infrared observations of thermal dust emission from supernova 1987A. *Nature* **340**, 697–699 (1989).
- Dunne, L., Eales, S., Ivison, R., Morgan, H. & Edmunds, M. Type II supernovae as a significant source of interstellar dust. *Nature* **424**, 285–287 (2003).
- Morgan, H. L., Dunne, L., Eales, S. A., Ivison, R. J. & Edmunds, M. G. Cold dust in Kepler's supernova remnant. *Astrophys. J.* **597**, L33–L36 (2003).
- Richards, G. T. *et al.* Red and reddened quasars in the Sloan Digital Sky Survey. *Astron. J.* **126**, 1131–1147 (2003).
- Reichard, T. A. *et al.* Continuum and emission-line properties of broad absorption line quasars. *Astron. J.* **126**, 2594–2607 (2003).
- Hopkins, P. *et al.* *Astrophys. J.* (in the press); preprint at (<http://arXiv.org/astro-ph/0406293>) (2004).
- Gaskell, C. M., Goosmann, R. W., Antonucci, R. R. J. & Whyson, D. H. The nuclear reddening curve for active galactic nuclei and the shape of the infra-red to X-ray spectral energy distribution. *Astrophys. J.* (submitted); preprint at (<http://arXiv.org/astro-ph/0309595>) (2003).
- Maiolino, R. *et al.* Dust in active nuclei. I. Evidence for “anomalous” properties. *Astron. Astrophys.* **365**, 28–36 (2001).
- Maiolino, R. *et al.* Extreme gas properties in the most distant quasars. *Astron. Astrophys.* **420**, 889–897 (2004).
- Baffa, C. *et al.* NICS: The TNG Near Infrared Camera Spectrometer. *Astron. Astrophys.* **378**, 722–728 (2001).
- Maiolino, R., Juarez, Y., Mujica, R., Nagar, N. & Oliva, E. Early star formation traced by the highest-redshift quasars. *Astrophys. J.* **596**, L155–L158 (2003).
- Hirashita, H., Hunt, L. K. & Ferrara, A. Dust and hydrogen molecules in the extremely metal-poor dwarf galaxy SBS 0335–052. *Mon. Not. R. Astron. Soc.* **330**, L19–L23 (2002).
- Woosley, S. E. & Weaver, T. A. The evolution and explosion of massive stars. II. Explosive hydrodynamics and nucleosynthesis. *Astrophys. J. Suppl.* **101**, 181–235 (1995).
- Larson, R. B. Early star formation and the evolution of the stellar initial mass function in galaxies. *Mon. Not. R. Astron. Soc.* **301**, 569–581 (1998).
- Zubko, V. G., Mennella, V., Colangeli, L. & Bussoletti, E. Optical constants of cosmic carbon analogue grains. I. Simulation of clustering by a modified continuous distribution of ellipsoids. *Mon. Not. R. Astron. Soc.* **282**, 1321–1329 (1996).
- Scott, A. & Duley, W. W. Ultraviolet and infrared refractive indices of amorphous silicates. *Astrophys. J. Suppl.* **105**, 401–405 (1996).
- Mukai, T. in *Evolution of Interstellar Dust and Related Topics* (eds Bonetti, A., Greenberg, J. M. & Aiello, S.) 397–446 (Elsevier Science, New York, 1989).

Acknowledgements All authors have contributed equally to this paper. We thank J. Brucato for providing the optical constant of ACAR grains and we thank M. Walmsley for comments. This work was partially supported by the Italian Ministry of Research (MIUR) and by the National Institute for Astrophysics (INAF).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.M. (maiolino@arcetri.astro.it).

Monoenergetic beams of relativistic electrons from intense laser–plasma interactions

S. P. D. Mangles¹, C. D. Murphy^{1,2}, Z. Najmudin¹, A. G. R. Thomas¹, J. L. Collier², A. E. Dangor¹, E. J. Divall², P. S. Foster², J. G. Gallacher³, C. J. Hooker², D. A. Jaroszynski³, A. J. Langley², W. B. Mori⁴, P. A. Norreys², F. S. Tsung⁴, R. Viskup³, B. R. Walton¹ & K. Krushelnick¹

¹The Blackett Laboratory, Imperial College London, London SW7 2AZ, UK

²Central Laser Facility, Rutherford Appleton Laboratory, Chilton, Didcot, Oxon, OX11 0QX, UK

³Department of Physics, University of Strathclyde, Glasgow G4 0NG, UK

⁴Department of Physics and Astronomy, UCLA, Los Angeles, California 90095, USA

High-power lasers that fit into a university-scale laboratory¹ can now reach focused intensities of more than 10^{19} W cm^{−2} at high repetition rates. Such lasers are capable of producing beams of energetic electrons^{2–11}, protons¹² and γ -rays¹³. Relativistic electrons are generated through the breaking^{9,10,14} of large-amplitude relativistic plasma waves created in the wake of the laser pulse as it propagates through a plasma, or through a direct interaction between the laser field and the electrons in the plasma¹⁵. However, the electron beams produced from previous laser–plasma experiments have a large energy spread^{6,7,9,14}, limiting their use for potential applications. Here we report high-resolution energy measurements of the electron beams produced from intense laser–plasma interactions, showing that—under particular plasma conditions—it is possible to generate beams of relativistic electrons with low divergence and a small energy spread (less than three per cent). The monoenergetic features were observed in the electron energy spectrum for plasma densities just above a threshold required for breaking of the plasma wave. These features were observed consistently in the electron spectrum, although the energy of the beam was observed to vary from shot to shot. If the issue of energy reproducibility can be addressed, it should be possible to generate ultrashort monoenergetic electron bunches of tunable energy, holding great promise for the future development of ‘table-top’ particle accelerators.

Recently, electrons with energies greater than 200 MeV have been observed using ultrashort pulse lasers (the forced laser wakefield regime)¹⁴, while electrons having an energy up to 350 MeV have been observed from laser plasma interactions using petawatt (10^{15} W)-scale lasers¹⁶. These energies were obtained from acceleration distances of a millimetre or less. However, the electron beams produced from laser–plasma experiments have always previously been observed to have a large energy spread (that is, $\Delta E/E = 100\%$). Although schemes have been proposed to produce monoenergetic beams by injecting electrons into a particular phase of a plasma wave using complex multiple laser beam geometries^{17,18} or external electron beams¹⁹, no success has yet been reported. If reproducible monoenergetic beams could be developed, then table-top narrowband femtosecond X-ray sources and free-electron lasers could become a reality, which could potentially lead to significant advances in both medicine and material science. It may also be possible to use the electron bunches generated in this way for injection into conventional radio frequency (RF) accelerators or into subsequent plasma acceleration stages.

The experiment used the Astra laser at the Rutherford Appleton Laboratory to focus ultrashort ($\tau = 40$ fs, 0.5 J) pulses onto a supersonic helium gas-jet (see Fig. 1). The plasma density was varied from $n_e = 3 \times 10^{18}$ cm^{−3} to $n_e = 5 \times 10^{19}$ cm^{−3} by varying the gas-jet pressure. In this density range, the wavelength

of relativistic plasma waves produced ($\lambda_p = 2\pi c/\omega_{pe}$) is between 0.33 and 2 times the laser pulse length ($c\tau = 12\ \mu\text{m}$), where $\omega_{pe} = (n_e e^2/m\epsilon_0)^{1/2}$ is the plasma frequency. Over this range of densities, plasma waves are driven by the ponderomotive force of the laser, which is proportional to the intensity gradient of the pulse. The ponderomotive force predominantly pushes electrons forwards (there is also a radial push). Because the ions are much heavier than the electrons and so do not respond to the ponderomotive force, the electrons are dragged back towards their original position by the space charge field. As they overshoot, a plasma oscillation is formed. As the laser beam moves forward through the plasma this sets up a plasma wave travelling in its wake which has a phase velocity equal to the group velocity of the laser pulse in the plasma^{20–22}. For low laser intensities, this process is most efficient for pulse lengths where the pulse is shorter than the plasma wavelength. At higher intensities, nonlinear modification of the laser pulse by the plasma wave can efficiently drive plasma waves^{14,23}, even when this condition is not initially met.

Once established, the plasma wave can then grow until wave-breaking occurs^{9,24}. This is where, at very large plasma wave amplitude, the wave motion becomes so nonlinear that wave energy is transferred directly into particle energy. Wave-breaking is not always catastrophic and a proportion of the electrons in the wave can break from the wave, reducing its amplitude, while maintaining the wave structure. Thus this population of injected electrons can continue to interact with the wave and gain further energy.

In our experiment, the electron energy spectrum was measured using an on-axis magnetic spectrometer. A high-resolution image-plate detector (Fuji BAS1800II) was used to obtain the electron spectrum. The electrons were also simultaneously measured using a lower-resolution array of diodes situated behind the image plate in order to calibrate the image plate response. This set-up was able to measure the spectrum over a wide energy range in a single shot.

Electron acceleration was observed over a range of electron densities. With the plasma density below $7 \times 10^{18}\ \text{cm}^{-3}$, no energetic electrons were observed (this corresponds to $\lambda_p = c\tau$).

Increasing the density produced a sudden change, with the detection of energetic electrons up to 100 MeV. Measurements of the beam divergence using radiochromic film detectors show that the full-width at half-maximum (FWHM) of the electron beam was less than 5° . However, the most interesting aspect of these spectra is that, in this regime, the electron energy spectra were exceptionally non-maxwellian. Indeed they generally consisted of one or more narrow spiky features, each having an energy bandwidth of less than 20% (see Fig. 2). As the density was increased further, the peak energy of the electrons was observed to decrease and the spectra began to assume a broad maxwellian shape, as reported in previous experiments (see Fig. 2).

The difference observed in these spectra can be attributed to the timing of the injection of electrons into the relativistic plasma wave. Evidently wave-breaking places the electrons to be accelerated at a precise phase within the plasma wave. In this way, all the electrons experience an almost identical acceleration gradient. As the wake-field is several plasma wavelengths in duration, with its amplitude decreasing away from the laser pulse, successive plasma periods can accelerate trapped electron bunches to different energies, producing the multiple spikes in the spectrum.

With careful control of the plasma density and at a higher laser power, the monoenergetic structure was even clearer (Fig. 3); typically, only one very narrow single peak in the spectrum was observed. In this case it is likely that only the first plasma oscillation is driven to breaking point. Note that the total number of electrons in the peak in Fig. 3 is estimated to be about 1.4×10^8 or $\sim 22\ \text{pC}$, with a FWHM energy spread of $<3\%$. Under the same experimental conditions the spectrum consistently showed narrow energy spread, but with variation in the energy of the peak. For shots with the lowest energy spread ($<10\%$) the beam energy varied between 50 and 80 MeV. This is almost certainly due to shot-to-shot variations in the laser parameters.

For these monoenergetic beams to propagate out of the plasma, the electron bunches cannot be dephased (that is, outrun the plasma wave), as then they would be decelerated by the front of the plasma wave. This means that the sum of the length after which wave-breaking occurs and the dephasing length needs to be greater than the interaction length. Indeed, the plasma density regime in which these narrow-energy-spread beams were observed is that where the dephasing length was longer than the observed inter-

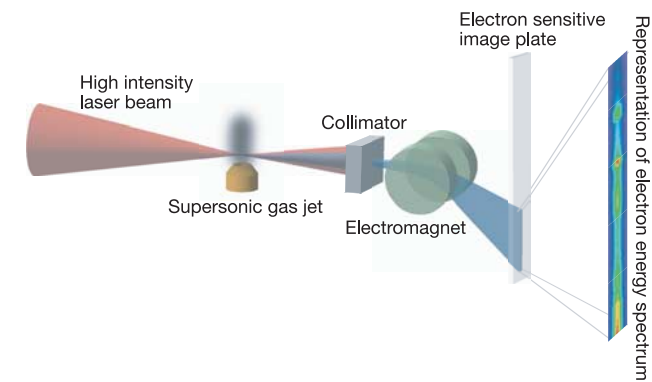


Figure 1 Experimental set-up. The experiment used the high-power titanium:sapphire laser system at the Rutherford Appleton Laboratory (Astra). The laser pulses ($\lambda = 800\ \text{nm}$, $\tau = 40\ \text{fs}$ with energy approximately $0.5\ \text{J}$ on target) were focused with an $f/16.7$ off-axis parabolic mirror onto the edge of a 2-mm-long supersonic jet of helium gas to produce peak intensities up to $2.5 \times 10^{18}\ \text{W cm}^{-2}$. The Astra laser has typical shot-to-shot reproducibility for high-power lasers, with variations in pulse energy $\pm 5\%$, pulse length $\pm 12\%$ and focal spot size $\pm 11\%$, while the focal spot can move by up to a spot diameter ($\sim 25\ \mu\text{m}$ in this case). The electron density (n_e) as a function of backing pressure on the gas jet was determined by measuring the frequency shift ($\Delta\omega = \omega_{pe}$, where ω_{pe} is the electron plasma frequency) of satellites generated by forward Raman scattering in the transmitted laser spectrum⁵. The plasma density was observed to vary linearly with backing pressure within the range $n_e = 3 \times 10^{18}\ \text{cm}^{-3} - 5 \times 10^{19}\ \text{cm}^{-3}$. Electron spectra are measured using an on-axis magnetic spectrometer. Other diagnostics used included transverse imaging of the interaction, and radiochromic film stacks to measure the divergence and total number of accelerated electrons.

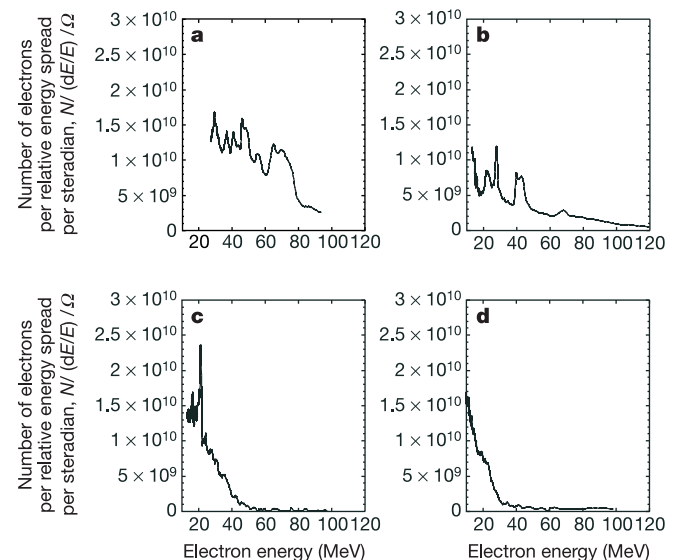


Figure 2 Measured electron spectra at various densities. Laser parameters: $E \approx 350\ \text{mJ}$, $\tau \approx 40\ \text{fs}$, $I \approx 1.5 \times 10^{18}\ \text{W cm}^{-2}$. Densities (n_e , in units of $10^{19}\ \text{cm}^{-3}$): **a**, 1.6; **b**, 1.8; **c**, 3; and **d**, 5.

action length (see Fig. 4). In contrast, at higher densities the dephasing distance is shorter than the interaction distance, and so a quasi-maxwellian distribution of electrons emerges from the plasma (Fig. 2d).

This acceleration mechanism described is supported by particle-in-cell simulations of the interaction, performed using the code OSIRIS²⁵ on an eight G5 node "Applecluster" at Imperial College London. The simulations were performed over the range of our experimental parameters, and in 2D3V (two spatial but three momentum and field dimensions.) As previously noted¹⁴, 2D3V simulations can underestimate the maximum electron energies,

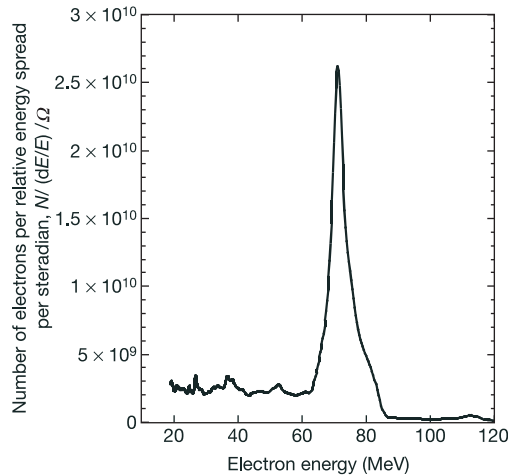


Figure 3 Measured electron spectrum at a density of $2 \times 10^{19} \text{ cm}^{-3}$. Laser parameters: $E = 500 \text{ mJ}$, $\tau = 40 \text{ fs}$, $I \approx 2.5 \times 10^{18} \text{ W cm}^{-2}$. The energy spread is $\pm 3\%$. The energy of this monoenergetic beam fluctuated by $\sim 30\%$, owing to variations in the laser parameters.

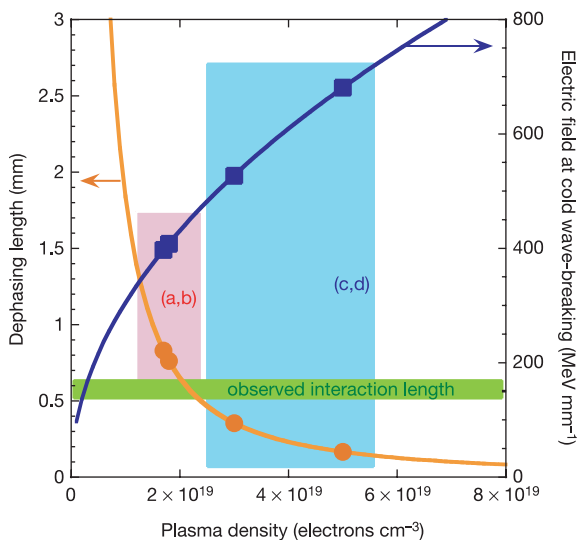


Figure 4 Plot of dephasing length and cold wave-breaking amplitude versus plasma density. Simulations show that the dephasing length in a nonlinear plasma wave remains close to the linear value, $L_d \approx 2\pi c \omega^2 / \omega_{pe}^3$, owing to competition between the nonlinearly increasing plasma wavelength, and the decrease in laser pulse group velocity due to photon deceleration. The dephasing lengths (circles) and wave-breaking amplitudes (squares) corresponding to the spectra shown in Fig. 2 are indicated; those in the red shaded region correspond to the spectra that exhibited monoenergetic features, and those in the blue shaded region correspond to the spectra that exhibited maxwellian energy distributions. The green line indicates the interaction length observed using transverse imaging diagnostics.

owing to reductions in the degrees of freedom for self-focusing and plasma wave growth. However, they do accurately describe the phenomenology of the interaction.

As in the experiments, the simulations show that for plasma densities for which the plasma wavelength is greater than the pulse length ($\lambda_p \geq c\tau$), a plasma wave is generated, but there is no wave-breaking. At these low densities the forced laser wakefield mechanism¹⁴ is ineffective. But at densities slightly above this threshold, a noticeable change occurs in the interaction. The generation of the plasma wave causes self-focusing of the laser pulse away from its leading edge, owing to the radial density profile of the plasma wave. It is noted that for short pulses relativistic self-focusing is ineffective for the front of the pulse²⁶. The laser pulse becomes shaped like a cone, tapered towards the rear, with a length close to λ_p . This causes a feedback mechanism, where the increasing laser intensity towards the back causes the plasma wave amplitude to grow, which can further focus the laser pulse. As the plasma wave reaches large amplitude the longitudinal motion of the electrons in the wave becomes relativistic, which leads to a lengthening of the plasma wavelength.

Crucially, as the laser pulse length is now less than the plasma wavelength, plasma electrons can stream into the plasma wave transversely behind the laser pulse, where previously they were excluded by the laser's ponderomotive force. Because the waveform is non-sinusoidal, a large number of electrons can be injected into a particular phase of the plasma wave and experience an accelerating force. This transverse breaking of the wave reduces the electric field strength of the plasma wave, thus preventing further injection and so ensuring an electron bunch localized in position and time. The transverse injection of electrons can explain why the plasma wave can break at amplitudes significantly less ($E \approx E_0$) than the one-dimensional cold wave-breaking limit, $E_{wb} = \sqrt{2}(\gamma_p - 1)^{1/2} E_0$, where γ_p is the Lorentz factor associated with the plasma wave ($\gamma_p \approx \omega/\omega_{pe}$) (ref. 24).

All of the electrons in this bunch then experience very similar acceleration as is demonstrated in Fig. 5, until they begin to outrun the steepened accelerating front of the plasma wave. If the length of

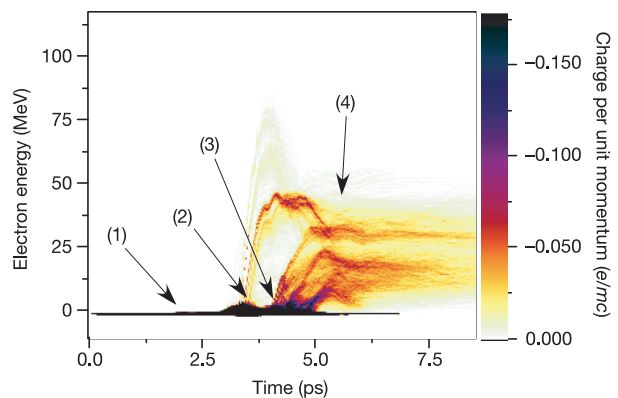


Figure 5 Evolution of the energy spectrum of the electrons (integrated over the two-dimensional simulation box) during a 1 mm interaction at a plasma density of $n_e = 2.1 \times 10^{19} \text{ cm}^{-3}$. The simulation space was $1,536 \times 1,024$ cells (16 cells per λ) with 4 electrons per cell. At the time indicated by the arrow (1) the pulse is self-focused and some relativistic electrons have appeared, but at quite low energies. The laser pulse front begins to steepen owing to the forced wakefield mechanism¹⁴ and this causes the wakefield amplitude to grow. At time (2), the plasma wavelength begins to increase relativistically, and at this point transverse wave-breaking takes place. This bunch experiences a uniform acceleration to high energy. At later time (3), further plasma oscillations, behind the initial one, also break transversely, resulting in multiple bunches of accelerated electrons. As they travel further, these electron bunches begin to dephase with respect to the plasma wave causing energy spread, just before they leave the plasma (4).

the plasma is sufficiently short, then one can extract this monoenergetic bunch, before further wave-breaking of plasma oscillations behind the first causes an increased 'dark-current' of lower-energy electrons (marked (3) in Fig. 5). Indeed, in some simulations this 'dark-current' is found to be greatly reduced if the trailing wake is sufficiently perturbed to prevent efficient trapping and acceleration.

In the simulation shown, the end of the plasma (at $t = 6.6$ ps) is sufficiently close that many of the monoenergetic features remain as they leave the plasma. If the plasma is too long, then the electrons begin to enter a decelerating part of the plasma wave and as a result experience longitudinal energy spread, as can be observed at (4) in Fig. 5. For higher densities, or longer plasmas, this dephasing is complete, and a maxwellian electron energy spectrum results. Such dephasing has been the case in most previous laser-based acceleration experiments, explaining why such features have not been seen before. However, with the right combination of laser pulse length ($c\tau \approx \lambda_p$), focal spot size, plasma density and interaction length, monoenergetic features such as that in Fig. 3 are the principal characteristic of the spectrum of relativistic electrons in both simulations and experiment. $\square$

Received 26 May; accepted 18 August 2004; doi:10.1038/nature02939.

1. Perry, M. D. & Mourou, G. Terawatt to petawatt subpicosecond lasers. *Science* **264**, 917–924 (1994).
2. Key, M. H. *et al.* Hot electron production and heating by hot electrons in fast ignitor research. *Phys. Plasmas* **5**, 1966–1972 (1998).
3. Tajima, T. & Dawson, J. Laser electron accelerator. *Phys. Rev. Lett.* **43**, 267–270 (1979).
4. Joshi, C. & Katsouleas, T. Plasma accelerators at the energy frontier and on tabletops. *Phys. Today* **56**, 47–53 (2003).
5. Esarey, E., Sprangle, P., Krall, J. & Ting, A. Overview of plasma-based accelerator concepts. *IEEE Trans. Plasma Sci.* **24**, 252–288 (1996).
6. Wagner, R., Chen, S. Y., Maksimchuk, A. & Umstadter, D. Electron acceleration by a laser wakefield in a relativistically self-guided channel. *Phys. Rev. Lett.* **78**, 3125–3128 (1997).
7. Ting, A. *et al.* Plasma wakefield generation and electron acceleration in a self-modulated laser wakefield accelerator experiment. *Phys. Plasmas* **4**, 1889–1899 (1997).
8. Tzeng, K. C. & Mori, W. B. Suppression of electron ponderomotive blowout and relativistic self-focusing by the occurrence of Raman scattering and plasma heating. *Phys. Rev. Lett.* **81**, 104–107 (1998).
9. Modena, A. *et al.* Electron acceleration from the breaking of relativistic plasma waves. *Nature* **377**, 606–608 (1995).
10. Santala, M. I. K. *et al.* Observation of a hot high-current electron beam from a self-modulated laser wakefield accelerator. *Phys. Rev. Lett.* **86**, 1227–1230 (2001).
11. Pukhov, A. & Meyer-ter-Vehn, J. Laser wake field acceleration: the highly non-linear broken-wave regime. *Appl. Phys. B* **74**, 355–361 (2002).
12. Clark, E. L. *et al.* Measurements of energetic proton transport through magnetized plasma from intense laser interactions with solids. *Phys. Rev. Lett.* **84**, 670–673 (2000).
13. Edwards, R. D. *et al.* Characterization of a gamma-ray source based on a laser-plasma accelerator with applications to radiography. *Appl. Phys. Lett.* **80**, 2129–2131 (2002).
14. Malka, V. *et al.* Electron acceleration by a wake field forced by an intense ultrashort laser pulse. *Science* **298**, 1596–1600 (2002).
15. Pukhov, A., Sheng, Z. M. & Meyer-ter-Vehn, J. Particle acceleration in relativistic laser channels. *Phys. Plasmas* **6**, 2847–2854 (1999).
16. Mangles, S. P. D. *et al.* Electron acceleration to 350 MeV due to the direct interaction of an ultra-intense laser pulse with an underdense plasma. *Phys. Rev. Lett.* (submitted).
17. Umstadter, D. Review of physics and applications of relativistic plasmas driven by ultra-intense lasers. *Phys. Plasmas* **8**, 1774–1785 (2001).
18. Umstadter, D., Kim, J. K. & Dodd, E. Laser injection of ultrashort electron pulses into wakefield plasma waves. *Phys. Rev. Lett.* **76**, 2073–2076 (1996).
19. Amiranoff, F. *et al.* Observation of laser wakefield acceleration of electrons. *Phys. Rev. Lett.* **81**, 995–998 (1998).
20. Sprangle, P., Esarey, E. & Ting, A. Nonlinear interaction of intense laser pulses in plasmas. *Phys. Rev. A* **41**, 4463–4467 (1990).
21. Bulanov, S. V., Kirsanov, V. I. & Sakharov, A. S. Excitation of ultrarelativistic plasma waves by pulse of electromagnetic radiation. *JETP Lett.* **50**, 198–201 (1989).
22. Berezhiani, V. I. & Murusidze, I. G. Relativistic wake-field generation by an intense laser-pulse in a plasma. *Phys. Lett. A* **148**, 338–340 (1990).
23. Sprangle, P., Esarey, E., Ting, A. & Joyce, G. Laser wakefield acceleration and relativistic optical guiding. *Appl. Phys. Lett.* **53**, 2146–2148 (1988).
24. Akhiezer, A. I. & Polovin, R. V. Theory of wave motion of an electron plasma. *JETP* **3**, 696–705 (1956).
25. Fonseca, R. A. *et al.* *Lecture Notes in Computer Science* Vol. 2329, III-342 (Springer, Heidelberg, 2002).
26. Faure, J. *et al.* Effects of pulse duration on self-focusing of ultra-short lasers in underdense plasmas. *Phys. Plasmas* **9**, 756–759 (2002).

Acknowledgements This work was supported by the UK EPSRC and RCUK. We thank the OSIRIS consortium (UCLA/IST Lisboa/USC) for the use of OSIRIS, and S. Karsch for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.M. (stuart.mangles@imperial.ac.uk).

High-quality electron beams from a laser wakefield accelerator using plasma-channel guiding

C. G. R. Geddes^{1,2}, Cs. Toth¹, J. van Tilborg^{1,3}, E. Esarey¹, C. B. Schroeder¹, D. Bruhwiler⁴, C. Nieter⁴, J. Cary^{4,5} & W. P. Leemans¹

¹Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA

²University of California, Berkeley, California 94720, USA

³Technische Universiteit Eindhoven, Postbus 513, 5600 MB Eindhoven, the Netherlands

⁴Tech-X Corporation, 5621 Arapahoe Ave. Suite A, Boulder, Colorado 80303, USA

⁵University of Colorado, Boulder, Colorado 80309, USA

Laser-driven accelerators, in which particles are accelerated by the electric field of a plasma wave (the wakefield) driven by an intense laser, have demonstrated accelerating electric fields of hundreds of GV m⁻¹ (refs 1–3). These fields are thousands of times greater than those achievable in conventional radio-frequency accelerators, spurring interest in laser accelerators^{4,5} as compact next-generation sources of energetic electrons and radiation. To date, however, acceleration distances have been severely limited by the lack of a controllable method for extending the propagation distance of the focused laser pulse. The ensuing short acceleration distance results in low-energy beams with 100 per cent electron energy spread^{1–3}, which limits potential applications. Here we demonstrate a laser accelerator that produces electron beams with an energy spread of a few per cent, low emittance and increased energy (more than 10⁹ electrons above 80 MeV). Our technique involves the use of a preformed plasma density channel to guide a relativistically intense laser, resulting in a longer propagation distance. The results open the way for compact and tunable high-brightness sources of electrons and radiation.

Previous laser driven plasma acceleration experiments have used a single intense (10¹⁸–10¹⁹ W cm⁻²) laser pulse propagating in the plasma formed when the front edge of the pulse ionized the gas plume emanating from a jet. The laser power was above the critical power for self-focusing and the laser pulse length exceeded the plasma period⁶. In this so-called self-modulated wakefield regime⁶, some self-guiding of the laser pulse occurs owing to relativistic modification of the plasma refractive index, but the laser pulse is highly unstable⁷. Large plasma waves, or wakes, can be driven by the radiation pressure of the intense laser, but the propagation length is limited to little more than the diffraction or Rayleigh length, Z_R (refs 5, 8). The best results have hence been obtained by increasing the laser spot size to increase Z_R , requiring ever greater laser power, but this approach has still been limited to distances of a few hundred micrometres (ref. 8). Plasma electrons were trapped and accelerated in the resulting high-amplitude plasma wave, and electron bunches with 100% energy spread and an exponentially small fraction of electrons at high energy were observed. For example, using a 32 TW

laser, bunches with more than 10^{10} electrons were produced but less than 1% were above 60 MeV (ref. 2). The beam divergence was typically greater than 50 mrad full-width at half maximum (FWHM). Whereas these beams are useful for some users^{9,10}, many applications, including high-energy physics and radiation sources, require monoenergetic beams (1% energy spread or less) with at least 10^9 electrons^{11–15}.

In the experiments reported here, a preformed plasma channel, with a radially symmetric density profile that has a minimum on the laser propagation axis, was used to guide the ultra-intense drive laser pulse, circumventing the limits imposed by diffraction¹⁴. For optimum acceleration, the accelerator length should be matched to the dephasing length, which is the distance it takes the accelerated electrons to outrun the plasma wave and slip into decelerating phase¹⁵. At high densities, where the laser pulse length is longer than the plasma wavelength, the laser can self-modulate, resulting in large plasma waves that trap particles from the bulk plasma. In a distance of the order of 2 mm, peak energy gain can be of the order of a few hundred MeV (ref. 2). With controlled injection^{16,17}, lower densities could be used in the future. The dephasing length and maximum energy increase as density decreases, so that accelerator lengths of several cm and energies up to several GeV per stage are then possible⁵. To reach high intensities with a typical 10 TW peak power laser system operating at 800 nm, the laser pulse must be focused to spot sizes near $8\text{ }\mu\text{m}$, giving $Z_R \approx 200\text{ }\mu\text{m}$ and necessitating guiding over many Z_R in order to reach the dephasing condition and high energies. A parabolic transverse density profile can be matched to guide a gaussian laser mode without distortion in the low intensity limit, and guiding has been previously demonstrated^{18–21} over many Z_R at input intensities below $2 \times 10^{17}\text{ W cm}^{-2}$. Such intensities are too low to trap and accelerate electrons. Here we report the first guiding of a relativistically intense laser (10^{19} W cm^{-2} input intensity) over a distance of $10\text{ }Z_R$, and the resultant production of high-quality electron beams. Much as in a RF accelerator where cavities provide confinement and field shaping for the RF power, the plasma channel guides the laser beam and transversely shapes the plasma wakes.

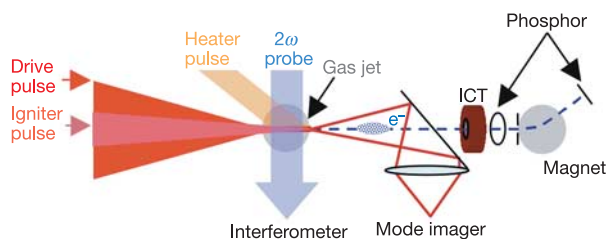


Figure 1 In the channel-guided laser wakefield accelerator, the plasma channel was formed in a supersonic hydrogen gas jet by two pulses fired 500 ps before the drive pulse. The supersonic gas jet was 2.4 mm long at an atomic density of $4.5 \times 10^{19}\text{ cm}^{-3}$. A cylindrical filament of plasma was ionized by an intense (60 fs, 15 mJ) igniter pulse, collinear with the pulse that drives the plasma wave and focused at #15 near the downstream edge of the gas jet. The plasma was subsequently heated to tens of eV by inverse bremsstrahlung, using a long (250 ps, 150 mJ) pulse incident from the side for efficient heating. The resulting hot plasma filament on axis expanded outward, driving a shock wave. This shock resulted in a density depletion on axis and a nearly parabolic transverse density profile which was tuned by adjusting the timing and energies of the beams. The plasma wave was driven by a 500 mJ pulse of 55 fs FWHM, focused at the upstream edge of the channel to an $8.5\text{ }\mu\text{m}$ FWHM spot by an $f/4$ off axis parabola giving an intensity of $1.1 \times 10^{19}\text{ W cm}^{-2}$. Propagation of the laser was monitored with a side interferometer (using a 2ω probe laser) and mode imager CCD. The electron beam accelerated by the plasma wave was analysed using an integrating current transformer (ICT), a phosphor screen, and a magnetic spectrometer (55° bend angle used for high resolution at energy $<92\text{ MeV}$, 5° for higher energies).

The present experiments used the multi arm L'OASIS Ti:Sapphire laser^{22,23}, operating at a wavelength of 810 nm with chirped pulse amplification²⁴, to form a guiding channel¹⁸ using a variation of the igniter-heater method¹⁹ and to drive the plasma wake (Fig. 1). By adjusting the energy and the timing of the guide formation pulses, the channel profile was matched to guide the drive pulse without distortion over $10\text{ }Z_R$ at input powers up to 4 TW ($7 \times 10^{18}\text{ W cm}^{-2}$). Guiding efficiency at 4 TW was 35%, a reduction of 30% compared to the low-power case (0.5 TW), suggesting that a significant amount of laser energy was depleted by excitation of a plasma wave. No electrons were accelerated at 4 TW, offering the possibility of using laser injection^{16,17} for the controlled trapping of electrons in this plasma wave without 'dark current'.

Electrons were trapped and accelerated using a 9 TW drive pulse, and optimal performance was found in channels detuned slightly from the low-power guiding condition, with 40% less rise in density over the spot diameter than a matched channel¹⁹. The contribution of relativistic self guiding present for the high-intensity pulse may be responsible for these changes. At 9 TW the mode image showed an intense output spot of $24\text{ }\mu\text{m}$ FWHM (Fig. 2b), slightly larger than the input spot but much smaller than the unguided pulse (Fig. 2d). Enlargement of the output spot compared to the input and leakage outside the guide appeared as the input intensity was increased well beyond the relativistic self focusing threshold, and are due to the inability of the channel to perfectly control the spot size in the presence of self guiding. Correlation of the mode imager and interferometer measurements showed that an intense guided mode was present only when the interferometer indicated that the laser was well confined to the channel (Fig. 2a). The axial plasma density was within 10% of $1.9 \times 10^{19}\text{ cm}^{-3}$ over the central 1.7 mm of the jet, and the drive laser pulse was thus a factor of two longer than the linear plasma period, that is, in the self-modulated regime⁶. This regime was chosen to allow comparison to unchannelled experiments, and also because the slower phase velocity of the wake at high plasma density allows trapping of background plasma electrons, yielding high-charge electron beams without a separate injector.

As shown in Fig. 3, the channel-guided accelerator produced

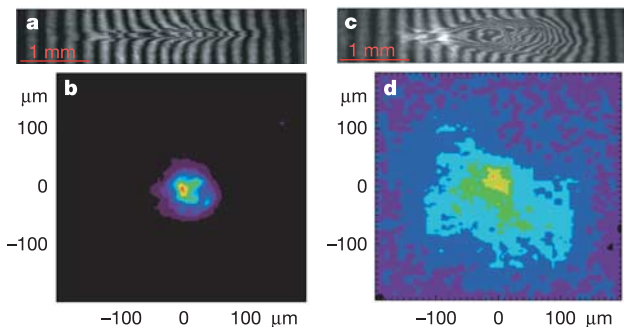


Figure 2 Laser propagation with and without channel. The plasma size after the propagation of the drive pulse shows the radial extent of ionization, indicating the extent of diffraction of the laser as it propagates through the plasma. The profile of the laser spot at the exit of the plasma was viewed using a solid mirror inserted into the beamline and an $f/10$ achromatic lens that imaged the laser spot directly onto a CCD (mode imager) camera providing $10\text{ }\mu\text{m}$ resolution. In the channel-guided accelerator, the plasma (a) was similar to the guiding channel indicating that the drive laser pulse was confined to the channel. The laser mode at the channel exit is a well defined spot of $24\text{ }\mu\text{m}$ FWHM containing 10% of the input energy (b). This indicates the effectiveness of the channel in maintaining the drive beam intensity and mode over many diffraction lengths. The spot is circular, confirming that the guiding channel is cylindrically symmetric. The reduction in energy from the input spot is due to a combination of leakage from the channel and depletion of the laser energy to excite the wake. When the channel was off, the interferometer showed blow out of the drive pulse after a few hundred micrometres (c), and the mode imager showed a diffuse transmitted spot (d).

high-charge electron beams with small energy spread at high energy, a unique feature that has not been observed in previous laser plasma acceleration experiments. To obtain the beam charge, the spectrometer phosphor screen has been calibrated against an integrating current transformer (ICT) and against radionuclide activation measurements⁹, both of which are consistent. For the data shown in Fig. 3, the $\pm 2\%$ energy spread of the peak centred at 86 MeV is essentially limited by the spectrometer resolution, so that the beam may in fact have narrower energy spread. The divergence of the bunch at 86 MeV was half that of the integrated beam observed on the phosphor before the magnet, consistent with previous experiments that have shown that higher-energy electrons were more collimated². The bunch shows a contrast ratio greater than 10:1 above a broad distribution of charge extending on either side. Space charge effects should be minimal for this relativistic beam, so assuming ballistic propagation from a source the size of the laser spot in the channel (8.5 μm at entrance, 24 μm at exit), an upper limit for the emittance can be obtained. Femtosecond bunches have hence been produced²⁵ containing 2×10^9 electrons with geometric and normalized r.m.s. emittances ϵ_x below 0.05–0.01 π mm mrad and below 1–2 π mm mrad r.m.s., respectively. The peak current is of the order of 10 kA with emittance comparable to the best state of the art RF facilities¹¹. Bunches with energy up to 150 MeV have been observed on separate shots (see below).

Two-dimensional simulations using the particle in cell (PIC) code VORPAL²⁶ with parameters close to the experiment indicate that the high-quality electron bunches observed may be formed by a combination of pulse evolution, beam loading, and dephasing. As the drive laser pulse propagates through the plasma it self modulates, driving an intense plasma wake that traps electrons. The initial bunch of trapped electrons induces a secondary wake which interferes with the primary wake, reducing its amplitude²⁷. If the drive laser pulse energy is just above the threshold for trapping, this beam loading effect suppresses further injection, creating an electron bunch isolated in phase space. The trapped electrons are then accelerated until they outrun (dephase from) the wake, at which point they are concentrated in phase and energy, forming a high-quality bunch with low energy spread (Fig. 4). Matching accelerator length and dephasing length to obtain high-quality bunches with the parameters (jet length and Z_R) of this experiment required maintaining the intensity of the laser over many Z_R using a guiding

channel. This dephasing condition can alternatively be met by a short plasma at high density (see below) or by using a larger laser spot size to extend Z_R , but these alternatives are less efficient since high density lowers peak energy while large spot size requires many times greater laser power. Fluid²⁸ and other PIC²⁹ simulations have also observed that longer acceleration length results in narrow energy spread. Only one accelerating period of the plasma wave contributes to the high-energy beam due to beam loading in these simulations, so that the bunch length is less than a plasma period; for the experimental parameters the bunch length is near 10 fs FWHM.

The quality of optical guiding as well as the pointing, quality and charge of the electron beams at high energy fluctuated from shot to shot, probably caused by laser pointing jitter that changes the overlap between the wake drive pulse and the channel formation pulses (and hence guide quality and incoupling) as well as laser power fluctuations. Beams with 3×10^9 electrons have been observed at similar energies (78 ± 3 MeV FWHM), and electrons were observed up to the limit of our 55° high-resolution spectrometer (92 MeV). Using a separate phosphor screen placed after the magnet at a 5° angle to the beam, we observed bunches at energies up to 150 MeV, but this diagnostic does not allow the fine resolution required to resolve energy spread. Structure in the energy spectrum has been seen for electrons with energy as low as 15 MeV. Below 15 MeV there is an essentially continuous distribution, and total beam charge was 1.7×10^{10} electrons as measured by the ICT, subtending $f/8$. Using a bend magnet, the low-energy contribution can be separated, leaving a high-energy, high-quality beam with a few times 10^9 electrons. Consistent control of guiding and electron injection^{16,17} in order to stabilize these beams are among the next challenges for laser accelerators.

To provide a baseline for evaluating the effects of guiding, the accelerator was operated with the same gas jet and laser but without the guiding channel. As seen in Fig. 2c and d, the laser pulse diffracted strongly, limiting the acceleration length. The electron beam had a total charge of 1.5×10^{10} electrons, as measured by the ICT, with divergence near 50 mrad FWHM. The electron energy spectrum is described by a two-temperature Boltzmann distribution characterized by a 2.6 MeV temperature below 10 MeV and an 8 MeV temperature above 10 MeV. Structure in the energy spectrum did occur occasionally in the tail of the distribution above 15 MeV containing $<2\%$ of the charge, consistent with a beam accelerated over a short distance and with previous exper-

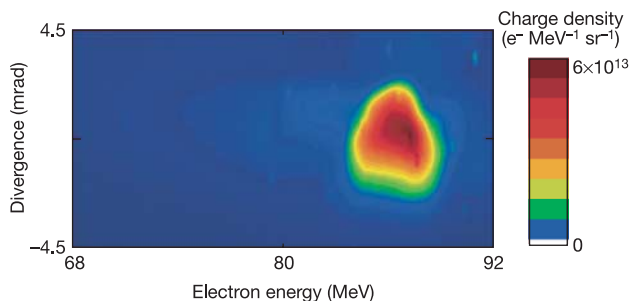


Figure 3 Single-shot electron beam spectrum and divergence of the channel-guided accelerator, showing a bunch containing 2×10^9 electrons in a narrow distribution at 86 ± 1.8 MeV and 3 mrad divergence FWHM with contrast $>10:1$ above background. This distribution is qualitatively different from the exponential distribution obtained in past (unchannelled) laser acceleration experiments. The magnetic spectrometer consists of a slit 82 cm from the gas jet, a bend of 55° in a dipole magnet to provide dispersion, and a phosphor screen (LANEX Fast backed by an aluminium foil to reject laser light) imaged by a CCD camera. Single-shot energy range is $\pm 15\%$ about a central value selectable from 1 to 80 MeV, and resolution is $dE/E = \pm 2\%$. The vertical beam size is obtained in the undispersed direction, allowing the simultaneous determination of (vertical) divergence and energy. Since electron beams observed on the phosphor before the magnet were typically round in shape, this vertical divergence measurement is representative.

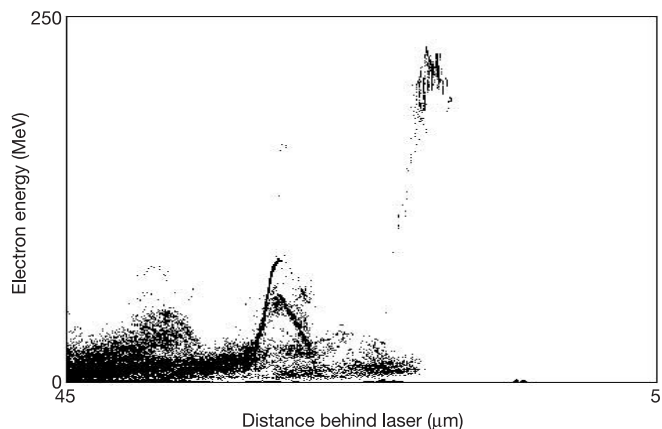


Figure 4 Particle in cell simulations, here displaying the phase space of the electrons, show an energy distribution similar to that in the experiments. A high-quality electron bunch is formed when the acceleration length is matched to the dephasing length, and when the laser strength is such that beam loading is sufficiently strong to turn off injection after the initial bunch of electrons is loaded. The peak energy observed in the simulations is 200 MeV, close to the experimental result.

iments^{1,2,9}. No electrons were observed above 40 MeV on the spectrometer phosphor screen (detection threshold 10^7 electrons). Plasma density was independently optimized for the unchannelled accelerator, and highest charge and electron energies were obtained at $4 \times 10^{19} \text{ cm}^{-3}$. The high density allowed acceleration before the laser diffracted since self modulation and dephasing both occur more quickly at high density, but this also reduced the peak energy compared to the channelled case. Operating the unchannelled accelerator at the density used for the channelled accelerator ($2 \times 10^{19} \text{ cm}^{-3}$) produced low-charge low-energy beams, since the intensity of the drive beam was not maintained for sufficient distance to allow acceleration at this density without channelling. Using a 600- μm -long plasma at $4 \times 10^{19} \text{ cm}^{-3}$, which is close to the dephasing length observed in PIC simulations at this density, the unchannelled accelerator produced the same peak energy as in the 2 mm plasma, but with more structure in the spectrum. These results confirm that matching accelerator length to dephasing length is critical for structuring the spectrum, and that extending the length (for instance, using a channel at lower density) results in higher energies.

To differentiate the effects of channelling from pre-ionization, the igniter pulse was fired 80 ps before the drive pulse. The plasma does not expand significantly over 80 ps, so there was no shock wave and the transverse density profile was flat and had no guiding properties. We observed no difference between the drive pulse only and pre-ionized cases, indicating that channelling and not pre-ionization was responsible for differences in the electron beams described above.

The beams from channel-guided accelerators such as those described here, with a few times 10^9 electrons in per cent level energy spread and mrad divergence, as well as intrinsic synchronization to the laser beam, open up a new class of experiments with laser accelerators. The channelling technology offers the ability to control the laser beam propagation much as a copper structure provides guiding and field shaping to RF accelerators. The investment of power in channel formation is less than 5% of the drive pulse power (20% of the energy), yet the spectral density of these beams near 80 MeV is at least a factor of 200 above previous unchannelled experiments using several times the laser power, and peak energy observed is comparable². The narrow energy spread of the channel-produced beams is consistent with simulations, which also indicate that the bunch length is near 10 fs. The channel-guided laser accelerator technique will hence allow efficient generation of femtosecond X-rays¹⁰, coherent THz and infrared radiation^{13,30}, and is an essential step towards the development of compact multistage electron accelerators with ultrafast bunches and with focusability and luminosity competitive with state of the art RF accelerators. □

Received 4 June; accepted 29 July 2004; doi:10.1038/nature02900.

- Modena, A. *et al.* Electron acceleration from the breaking of relativistic plasma waves. *Nature* **377**, 606–608 (1995).
- Malka, V. *et al.* Electron acceleration by a wake field forced by an intense ultrashort laser pulse. *Science* **298**, 1596–1600 (2002).
- Leemans, W. P. *et al.* Electron-yield enhancement in a laser-wakefield accelerator driven by asymmetric laser pulses. *Phys. Rev. Lett.* **89**, 174802 (2002).
- Tajima, T. & Dawson, J. M. Laser electron accelerator. *Phys. Rev. Lett.* **43**, 267–270 (1979).
- Esarey, E., Sprangle, P., Krall, J. & Ting, A. Overview of plasma-based accelerator concepts. *IEEE Trans. Plasma Sci.* **24**, 252–288 (1996).
- Esarey, E., Krall, J. & Sprangle, P. Envelope analysis of intense laser pulse self-modulation in plasmas. *Phys. Rev. Lett.* **72**, 2887–2890 (1994).
- Esarey, E., Sprangle, P., Krall, J. & Ting, A. Self-focusing and guiding of short laser pulses in ionizing gases and plasmas. *IEEE J. Quant. Electron.* **33**, 1879–1914 (1997).
- Najmudin, Z. *et al.* Self-modulated wakefield and forced laser wakefield acceleration of electrons. *Phys. Plasmas* **10**, 2071–2077 (2003).
- Leemans, W. P. *et al.* Gamma-neutron activation experiments using laser wakefield accelerators. *Phys. Plasmas* **8**, 2510–2516 (2001).
- Leemans, W. P. *et al.* Observation of terahertz emission from a laser-plasma accelerated electron bunch crossing a plasma-vacuum boundary. *Phys. Rev. Lett.* **91**, 074802 (2003).
- Catravas, P., Esarey, E. & Leemans, W. P. Femtosecond x-rays from Thomson scattering using laser wakefield accelerators. *Meas. Sci. Technol.* **12**, 1828–1834 (2001).
- Wang, X. J., Qiu, X. & Ben-Zvi, I. Experimental observation of high-brightness microbunching in a photocathode RF electron gun. *Phys. Rev. E* **54**, R3121–R3124 (1996).

- Schoenlein, R. W. *et al.* Femtosecond X-ray pulses at 0.4 Å generated by 90° Thomson scattering — A tool for probing the structural dynamics of materials. *Science* **274**, 236–238 (1996).
- Sprangle, P., Esarey, E., Krall, J. & Joyce, G. Propagation and guiding of intense laser pulses in plasmas. *Phys. Rev. Lett.* **69**, 2200–2203 (1992).
- Leemans, W. P. *et al.* Plasma guiding and wakefield generation for second-generation experiments. *IEEE Trans. Plasma Sci.* **24**, 331–342 (1996).
- Umstadter, D., Kim, J. K. & Dodd, E. Laser injection of ultrashort electron pulses into wakefield plasma waves. *Phys. Rev. Lett.* **76**, 2073–2076 (1996).
- Esarey, E., Hubbard, R. F., Leemans, W. P., Ting, A. & Sprangle, P. Electron injection into plasma wake fields by colliding laser pulses. *Phys. Rev. Lett.* **79**, 2682–2685 (1997).
- Durfee, C. G. & Milchberg, H. M. Light pipe for high intensity laser pulses. *Phys. Rev. Lett.* **71**, 2409–2412 (1993).
- Volfbeyn, P., Esarey, E. & Leemans, W. P. Guiding of laser pulses in plasma channels created by the ignitor-heater technique. *Phys. Plasmas* **6**, 2269–2277 (1999).
- Kim, K. Y., Alexeev, I., Fan, J., Parra, E. & Milchberg, H. M. Plasma waveguides: Addition of end funnels and generation in clustered gases. *AIP Conf. Proc.* **647**, 646–653 (2002).
- Gaul, E. W. *et al.* Production and characterization of a fully ionized He plasma channel. *Appl. Phys. Lett.* **77**, 4112–4114 (2000).
- Toth, C. *et al.* Powerful, pulsed, THz radiation from laser accelerated relativistic electron bunches. *Proc. SPIE* **5448**, 491–504 (2004).
- Leemans, W. P. *et al.* Laser-driven plasma-based accelerators — Wakefield excitation, channel guiding, and laser triggered particle injection. *Phys. Plasmas* **5**, 1615–1623 (1998).
- Strickland, D. & Mourou, G. Compression of amplified chirped optical pulses. *Opt. Commun.* **56**, 219–221 (1985).
- Leemans, W. P. *et al.* Terahertz radiation from laser accelerated electron bunches. *Phys. Plasmas* **5**, 2899–2906 (2004).
- Nieter, C. & Cary, J. R. VORPAL: A versatile plasma simulation code. *J. Comput. Phys.* **196**, 448–473 (2004).
- Katsouleas, T., Wilks, S., Chen, S., Dawson, J. M. & Su, J. J. Beam loading in plasma accelerators. *Part. Accel.* **22**, 81–99 (1987).
- Reitsma, A. J. W. *et al.* Simulation of electron postacceleration in a two-stage laser wakefield accelerator. *Phys. Rev. ST Accel. Beams* **5**, 051301 (2002).
- Tsung, F. S. *et al.* Near GeV energy laser wakefield acceleration of self-injected electrons in a cm scale plasma channel. *Phys. Rev. Lett.* submitted.
- Saas, M. *et al.* A setup for ultrafast time-resolved x-ray absorption spectroscopy. *Rev. Sci. Instrum.* **75**, 24–30 (2004).

Acknowledgements This work was supported by the US Department of Energy and the National Science Foundation and used resources of the National Energy Research Scientific Computing Center at LBNL; C.G. was also supported by the Hertz Foundation. C.G. acknowledges his faculty advisor J. Wurtele. We appreciate contributions from G. Dugan, J. Faure, G. Fubiani, B. Nagler, K. Nakamura, N. Saleh, B. Shadwick, L. Archambault, M. Dickinson, S. Dimaggio, D. Syversrud, J. Wallig and N. Ybarrolaza.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.P.L. (wplemans@lbl.gov).

A laser-plasma accelerator producing monoenergetic electron beams

J. Faure¹, Y. Glinec¹, A. Pukhov², S. Kiselev², S. Gordienko², E. Lefebvre³, J.-P. Rousseau¹, F. Burgy¹ & V. Malka¹

¹Laboratoire d'Optique Appliquée, Ecole Polytechnique, ENSTA, CNRS, UMR 7639, 91761 Palaiseau, France

²Institut für Theoretische Physik, I, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

³Département de Physique Théorique et Appliquée, CEA/DAM Ile-de-France, 91680 Bruyères-le-Châtel, France

Particle accelerators are used in a wide variety of fields, ranging from medicine and biology to high-energy physics. The accelerating fields in conventional accelerators are limited to a few tens of MeV m⁻¹, owing to material breakdown at the walls of the structure. Thus, the production of energetic particle beams currently requires large-scale accelerators and expensive infrastructures. Laser-plasma accelerators¹ have been proposed as a next generation of compact accelerators because of the huge

electric fields they can sustain^{2–5} ($>100 \text{ GeV m}^{-1}$). However, it has been difficult to use them efficiently for applications because they have produced poor-quality particle beams with large energy spreads^{2–10}, owing to a randomization of electrons in phase space. Here we demonstrate that this randomization can be suppressed and that the quality of the electron beams can be dramatically enhanced. Within a length of 3 mm, the laser drives a plasma bubble¹¹ that traps and accelerates plasma electrons. The resulting electron beam is extremely collimated and quasi-monoenergetic, with a high charge of 0.5 nC at 170 MeV.

For most practical applications, high-quality particle beams with high spatial quality and monoenergetic energy distribution are required. A beam that does not satisfy these criteria would be hard to use, because it would be difficult to transport it and/or to focus it. In order to produce high-quality beams from plasma-based accelerators, two challenges have to be met: (1) the generation of an accelerating structure in the plasma, and (2) the trapping and acceleration of injected beam loads into the accelerating structure. In order to generate accelerating structures in the plasma, a focused ultraintense laser pulse is used to drive large-amplitude plasma waves. One possible method for achieving this is the laser wakefield, in which plasma waves are excited by the laser ponderomotive force. When the laser pulse length $c\tau$ (where c is the speed of light and τ is the pulse duration) is comparable to the plasma wavelength λ_p , the ponderomotive force, which is proportional to the gradient of the laser intensity, efficiently pushes plasma electrons out of the regions of strong laser field. Thus, electrons are separated from the ions, which do not move because of their higher mass. This creates the space charge field needed for particle acceleration, that is, the plasma wave.

The generation of intense accelerating fields in plasmas has been demonstrated in many experiments^{8,12,13}. Proof-of-principle experiments have shown the feasibility of externally injecting electrons from a conventional accelerator into the laser-driven plasma accelerating structure^{8–10}. However, the output beam quality has been poor: the electron energy distribution has had a 100% energy spread. Until now, the most widespread method for producing electron beams from plasmas has relied on the self-modulated laser wakefield accelerator^{14–16}. In this accelerator, the laser pulse is longer than the plasma wavelength. Under the influence of the self-modulation instability, its envelope modulates at the plasma

frequency and resonantly excites a plasma wave. When the plasma wave amplitude reaches the wave-breaking level, copious amounts of plasma background electrons are trapped in the plasma wave and accelerated. Numerous experiments have produced electron beams with nC charge and divergence varying from a few degrees to tens of degrees and maxwellian energy distributions^{2–4}. More recently, several groups^{5–7,17} have demonstrated that more compact lasers can be used to efficiently generate high-repetition-rate (10 Hz) electron sources, which could be used for applications. However, these beams still have very large energy spreads and a low number of electrons at high energy (typically $<1 \text{ pC}$ at $200 \pm 10 \text{ MeV}$). Previous experiments inherently produced poor-quality beams: wave-breaking occurred under the laser pulse envelope and the accelerated electrons were also under the influence of the ultra-intense laser field. Direct laser acceleration^{6,18} by transverse laser field caused the spatial beam quality to deteriorate, causing emittance growth.

Here we demonstrate the generation of high-quality electron beams from ultraintense laser–plasma acceleration. Extremely collimated beams with 10 mrad divergence and $0.5 \pm 0.2 \text{ nC}$ of charge at $170 \pm 20 \text{ MeV}$ have been produced. Contrary to all previous results obtained from laser–plasma accelerators, the electron energy distribution is quasi-monoenergetic. The number of high-energy electrons (170 MeV) is increased by at least three orders of magnitude with respect to previous work.

The experiment was performed by focusing a chirped pulse amplification laser^{19,20} onto a helium gas jet (Fig. 1). Figure 2a shows a picture of the electron beam when no magnetic field is applied. The electron beam is very well collimated, with a 10 mrad divergence (full-width at half-maximum, FWHM); to our knowl-

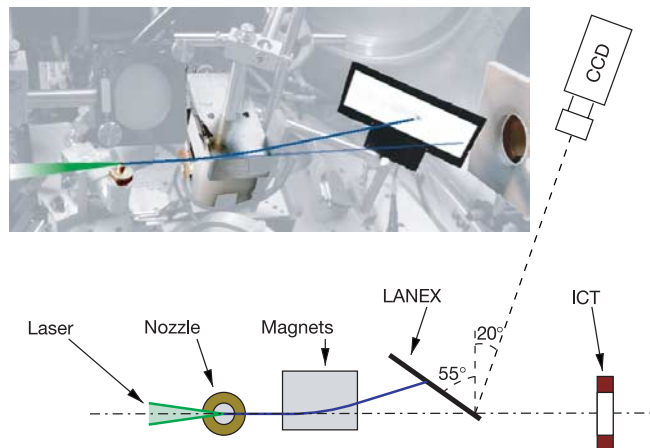


Figure 1 Experimental set-up. Top, picture of the experiment; bottom, diagram. An ultrashort and ultraintense laser pulse is focused onto a 3 mm supersonic gas jet and produces a highly collimated 170 MeV electron beam. LANEX is a phosphor screen; CCD, charge-coupled device camera; ICT, integrating current transformer.

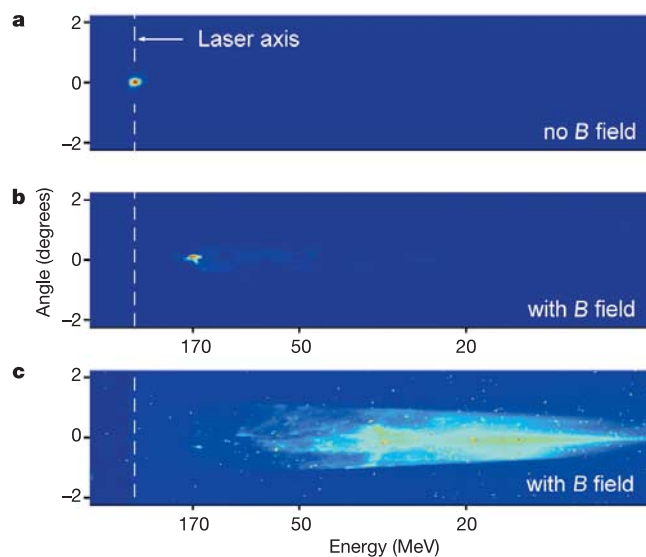


Figure 2 Raw images obtained on the LANEX screen. The vertical axis represents the beam angular divergence. When a magnetic field is applied, the horizontal axis represents electron energy. The white vertical dashed line is drawn at the intersection of the laser axis with the LANEX screen. **a**, Image of the electron beam spatial distribution obtained from the LANEX screen when no magnetic field (B) is applied. **b**, Image obtained when the magnetic field is applied, showing that the bulk of the beam is deviated and its position corresponds to 170 MeV electrons. The fact that the beam trajectory is displaced when a magnetic field is applied confirms that the signal on the LANEX screen corresponds to electrons and not to photons. **c**, Image obtained with a magnetic field and a higher plasma density ($n_e = 2 \times 10^{19} \text{ cm}^{-3}$). This electron beam has a much larger divergence and a 100% energy spread with few electrons above 100 MeV.

edge, this is the smallest divergence ever measured for a beam emerging from a plasma accelerator. Figure 2b shows the deviation of the beam when a magnetic field is applied. The image shows a narrow peak around 170 MeV, indicating efficient monoenergetic acceleration. For comparison, Fig. 2c shows an image obtained at higher electron density in the plasma ($n_e = 2 \times 10^{19} \text{ cm}^{-3}$). Here, electrons are randomly accelerated to all energies and the number of high-energy electrons is low. In addition, the beam divergence is much larger than in Fig. 2b. Figure 3 shows an electron spectrum after deconvolution. The distribution is clearly quasi-monoenergetic and peaks at 170 MeV, with a 24% energy spread (corresponding to the spectrometer resolution).

Finally, the charge contained in this beam can be inferred using an integrating current transformer: the whole beam contains $2 \pm 0.5 \text{ nC}$, and the charge at $170 \pm 20 \text{ MeV}$ is $0.5 \pm 0.2 \text{ nC}$. From the above, we can deduce that the electron beam energy was 100 mJ. Thus, the energy conversion from the laser to the electron beam was 10%.

Experimentally, this regime could be reached in a narrow range of parameters: stretching the pulse duration above 50 fs was sufficient to lose the peaked energy distribution. Similarly, when the electron density was increased from $6 \times 10^{18} \text{ cm}^{-3}$ to $7.5 \times 10^{18} \text{ cm}^{-3}$, the energy distribution became a broad plateau, similar to previous results⁵. Above 10^{19} cm^{-3} , the electron distribution was maxwellian-like with very few electrons accelerated at high energy. Below $6 \times 10^{18} \text{ cm}^{-3}$, the number of accelerated electrons decreased dramatically, although the distribution was still monoenergetic. The evolution of electron spectra with experimental parameters indicates that using laser pulses shorter than the plasma period is beneficial for high-quality and monoenergetic electron acceleration.

To reach a deeper understanding of the experiment, we have run three-dimensional (3D) particle-in-cell (PIC) simulations using the code Virtual Laser Plasma Laboratory²¹. The simulation results are shown in Fig. 4a–c. The simulation suggests that our experimental results can be explained by the following scenario. (1) At the beginning of the simulation, the laser pulse length ($9 \mu\text{m}$) is nearly resonant with the plasma wave ($\lambda_p = 13.6 \mu\text{m}$); but its diameter ($21 \mu\text{m} > \lambda_p$) is larger than the matched diameter. (2) As the pulse propagates in the plateau region of the gas jet, it self-focuses and undergoes longitudinal compression by plasma waves (Fig. 4a). This decreases the effective radius of the laser pulse and increases the

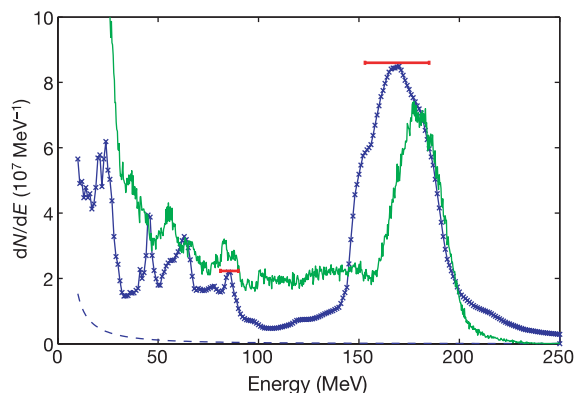


Figure 3 Experimental and simulated electron spectra. Blue line with crosses, electron spectrum corresponding to Fig. 2b, after deconvolution. Dashed line, estimation of the background level. Red horizontal error bars, resolution of the spectrometer. Green line, electron spectrum obtained from 3D PIC simulations. dN/dE is the number of electrons per MeV (E is the electron energy in MeV).

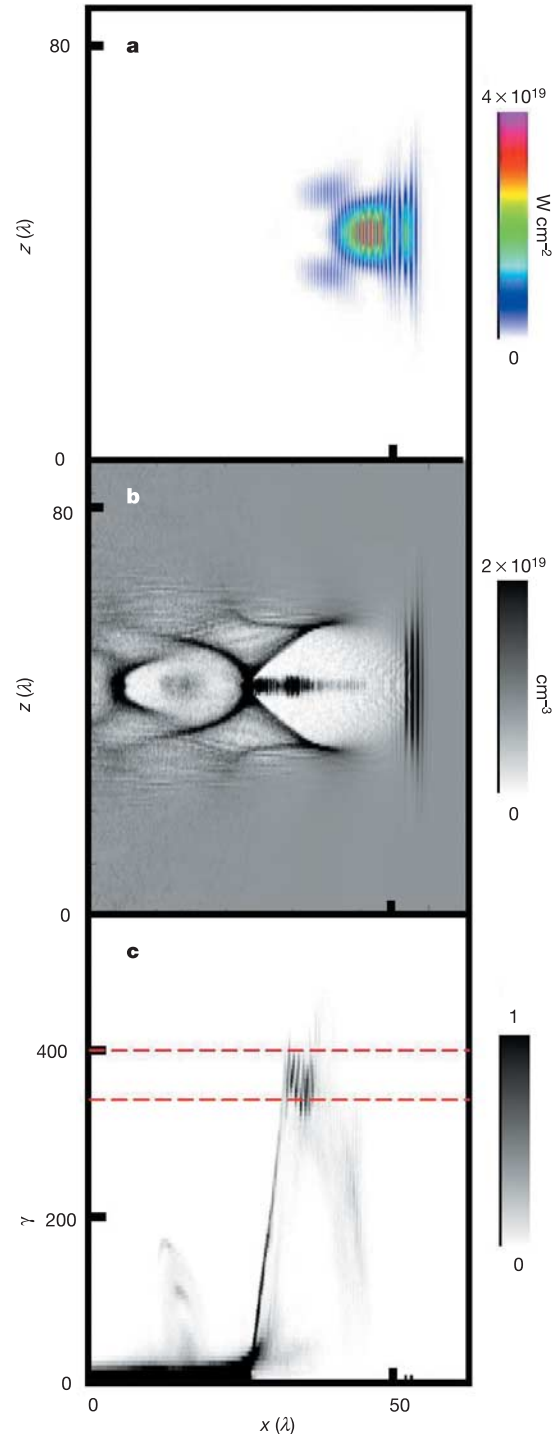


Figure 4 3D PIC simulation results. **a**, **b**, Distributions of laser intensity (**a**) and electron density (**b**) in the x - z plane, which is perpendicular to the polarization direction and passes through the laser axis. The laser pulse runs from left to right, and has propagated 2 mm in the plasma. The bubble structure is clearly visible. The laser pushes the electron fluid forward at the bubble head and creates a density compression there. Behind the laser we see the cavitated region with nearly zero electron density. The radially expelled electrons flow along the cavity boundary and collide at the X-point at the bubble base. Some electrons are trapped and accelerated in the bubble. The beam of accelerated electrons is seen as the black rod in **b**. These electrons are propagating behind the laser pulse (**a**) and are not disturbed by the laser field. **c**, Electron phase space density $f(x, \gamma)$ in arbitrary units. γ is the relativistic factor of the electron: $\gamma = (1 - v^2/c^2)^{-1/2}$, and v is the electron velocity. We see that the electrons have dephased and have self-bunched in the phase space around $\gamma \gg 350$. This self-bunching results in the mono-energetic peak in the energy spectrum (Fig. 3). The red horizontal dashed lines indicate the location of the mono-energetic peak in the phase space.

laser intensity by one order of magnitude. (3) This compressed laser pulse is now resonant with the plasma wave and it drives a highly nonlinear wakefield (Fig. 4b): the laser ponderomotive potential expels the plasma electrons radially and leaves a cavitating region behind (this is referred to as the 'cavitation' or 'blow-out' regime). In this regime, the 3D structure of the wakefield resembles a plasma bubble¹¹. (4) As the electron density at the walls of the bubble becomes large, wave-breaking occurs and electrons are injected and accelerated inside the bubble. (5) As the number of trapped electrons increases, the bubble elongates. Its effective group velocity decreases, and electrons start to dephase with respect to the accelerating field. This dephasing causes electron self-bunching in the phase space (Fig. 4c). This self-bunching results in the monoenergetic peak in the energy spectrum (Fig. 3).

Simulations also show that the quality of the electron beam is higher when trapped electrons do not interact with the laser field. If this were to occur, the laser field would cause the electrons to scatter in phase space, degrading the low divergence as well as the monoenergetic distribution. This argument could explain why higher-quality beams are obtained experimentally for shorter pulses and lower electron densities.

Figure 4a shows that the self-focused and compressed laser pulse stands in front of the trapped electrons (Fig. 4b), leaving them almost undisturbed^{5,11}. The electron energy spectrum obtained from the simulations is shown in Fig. 3: it peaks at 175 ± 25 MeV, in agreement with the experiment. The divergence of 10 mrad is also in agreement with experiments. Simulations also indicate that the electron bunch duration is less than 30 fs (here, the term 'bunch' refers to the fact that electrons are created in short bursts). Because the electron distribution is quasi-monoenergetic, the bunch will stay short upon propagation: considering a 24% energy spread at 170 MeV, we can show that the bunch stretches by only 50 fs m^{-1} as it propagates.

Another important point is the apparent robustness of the 'blow-out' regime. The initial laser parameters—for example, the focal spot radius and intensity—were far from the final values in the bubble (Fig. 4). Yet self-focusing led to compression of the laser pulse and to the formation of an electron cavity. The energy of 1 J for a 30 fs laser pulse, as used in the experiment, seems to be close to the threshold for this regime. Simulations¹¹ suggest that with more laser energy and shorter pulses, the blow-out regime and the formation of the bubble will lead to the acceleration of monoenergetic beams at higher energies and higher charges.

Our experimental results and 3D PIC simulations indicate that it is possible to generate a monoenergetic electron beam by carefully selecting laser and plasma parameters. The bunch duration (<50 fs), along with the present improvement in the charge (nC) and the quality of the electron beam (monoenergetic spectrum, low divergence), reinforce the relevance of plasma-based accelerators for many applications (such as high-resolution radiography for non-destructive material inspection, radiotherapy, ultrafast chemistry, radiobiology and material science). With the rapid progress of laser science, we expect that it will soon become possible to generate compact, monoenergetic and high-quality electron beams with a tunable energy range at a reasonable cost. Such a source would be perfectly adapted as an injector for future GeV laser-plasma accelerator schemes. It would also be relevant for generating ultrashort X-ray sources, using undulators or lasers via Thomson scattering. □

Methods

Laser

This new regime was reached by using the ultrashort and ultraintense laser pulse generated in a titanium-doped sapphire, chirped pulse amplification laser system. The laser pulse had a 33 ± 2 fs duration (FWHM), and contained 1 J of laser energy at central wavelength 820 nm. It was focused onto the edge of a 3-mm-long supersonic helium gas jet using a $f/18$

off-axis parabola. The diffraction-limited focal spot had a diameter of $r_0 = 21 \mu\text{m}$ at FWHM, producing a vacuum-focused laser intensity of $I = 3.2 \times 10^{18} \text{ W cm}^{-2}$, for which the corresponding normalized potential vector is $a_0 = eA/(mc^2) = 1.3$ (A is the laser vector potential, e and m are respectively the charge and mass of the electron). For these high laser intensities, the helium gas was fully ionized by the foot of the laser pulse and ionization did not play a role in the interaction.

Electron diagnostics

Electron detection was achieved using a LANEX phosphor screen, placed 25 cm after the gas jet. As electrons passed through the screen, energy was deposited and re-emitted into visible photons which were then imaged onto a 16-bit charge-coupled device (CCD) camera. For energy distribution measurements, a 0.45 T, 5-cm-long permanent magnet was inserted between the gas jet and the LANEX screen. The LANEX screen was protected by a 100- μm -thick aluminium foil in order to avoid direct exposure to the laser light. For deconvolution of the images obtained with the LANEX screen, electron deviation in the magnetic field has been considered as well as the electron stopping power inside the LANEX screen. The resolution (red error bar in Fig. 3) is limited by the electron beam spatial quality and by the dispersing power of the magnet. This gives a resolution of respectively 32 MeV and 12 MeV for 170 MeV and 100 MeV energies. Above 200 MeV, the resolution quickly degrades. The charge of the electron beam was measured using an integrating current transformer placed 30 cm behind the LANEX screen. It allowed us to measure the total charge of the beam when no magnetic field was applied, and the charge above 100 MeV when the magnetic field was applied.

PIC simulations

The simulation parameters corresponded to the optimal experimental case: the plasma electron density was $n_e = 6 \times 10^{18} \text{ cm}^{-3}$, the laser pulse duration was 30 fs and the initial laser spot size $21 \mu\text{m}$ FWHM. The laser pulse was assumed to be a perfect gaussian containing 1 J of energy. The plasma profile was chosen to fit the experimental density profile of the gas jet.

Received 5 July; accepted 25 August 2004; doi:10.1038/nature02963.

1. Tajima, T. & Dawson, J. M. Laser electron accelerator. *Phys. Rev. Lett.* **43**, 267–270 (1979).
2. Modena, A. et al. Electron acceleration from the breaking of relativistic plasma waves. *Nature* **337**, 606–608 (1995).
3. Umstadter, D., Chen, S.-Y., Maksimchuk, A., Mourou, G. & Wagner, R. Nonlinear optics in relativistic plasmas and laser wake field acceleration of electrons. *Science* **273**, 472–475 (1996).
4. Moore, C. I. et al. Electron trapping in self-modulated laser wakefields by Raman backscatter. *Phys. Rev. Lett.* **79**, 3909–3912 (1997).
5. Malka, V. et al. Electron acceleration by a wake field forced by an intense ultrashort laser pulse. *Science* **298**, 1596–1600 (2002).
6. Gahn, C. et al. Multi-MeV electron beam generation by direct laser acceleration in high-density plasma channels. *Phys. Rev. Lett.* **83**, 4772–4775 (1999).
7. Malka, V. et al. Characterization of electron beams produced by ultrashort (30 fs) laser pulses. *Phys. Plasmas* **8**, 2605–2608 (2001).
8. Kitagawa, Y. et al. Beat-wave excitation of plasma wave and observation of accelerated electrons. *Phys. Rev. Lett.* **68**, 48–51 (1992).
9. Everett, M. et al. Trapped electron acceleration by a laser-driven relativistic plasma wave. *Nature* **368**, 527–529 (1994).
10. Amiranoff, F. et al. Observation of laser wakefield acceleration of electrons. *Phys. Rev. Lett.* **81**, 995–998 (1998).
11. Pukhov, A. & Meyer-ter-Vehn, J. Laser wake field acceleration: the highly non-linear broken-wave regime. *Appl. Phys. B* **74**, 355–361 (2002).
12. Clayton, C. E., Joshi, C., Darrow, C. & Umstadter, D. Relativistic plasma-wave excitation by collinear optical mixing. *Phys. Rev. Lett.* **54**, 2343–2346 (1985).
13. Amiranoff, F. et al. Observation of modulational instability in Nd-laser beat-wave experiments. *Phys. Rev. Lett.* **68**, 3710–3713 (1992).
14. Andreev, N. E., Gorbunov, L. M., Kirsanov, V. I., Pogosova, A. A. & Ramazashvili, R. Resonant excitation of wakefields by a laser pulse in a plasma. *JETP Lett.* **55**, 571–574 (1992).
15. Sprangle, P., Esarey, E., Krall, J. & Joyce, G. Propagation and guiding of intense laser pulses in plasmas. *Phys. Rev. Lett.* **69**, 2200–2203 (1992).
16. Antonsen, T. M. & Mora, P. Self-focusing and Raman scattering of laser pulses in tenuous plasmas. *Phys. Rev. Lett.* **69**, 2204–2207 (1992).
17. Leemans, W. P. et al. Electron-yield enhancement in a laser-wakefield accelerator driven by asymmetric laser pulses. *Phys. Rev. Lett.* **89**, 174802 (2002).
18. Pukhov, A., Sheng, Z.-M. & Meyer-ter-Vehn, J. Particle acceleration in relativistic laser channels. *Phys. Plasmas* **6**, 2847–2854 (1999).
19. Strickland, D. & Mourou, G. Compression of amplified chirped optical pulses. *Opt. Commun.* **56**, 219–221 (1985).
20. Pittman, M. et al. Design and characterization of a near-diffraction-limited femtosecond 100-TW 10-Hz high-intensity laser system. *Appl. Phys. B* **74**, 529–535 (2002).
21. Pukhov, A. J. Three-dimensional electromagnetic relativistic particle-in-cell code VLPL (Virtual Laser Plasma Lab). *J. Plasma Phys.* **61**, 425–433 (1999).

Acknowledgements We acknowledge support from the European Community Research Infrastructure Activity under the FP6 "Structuring the European Research Area" programme (CARE) and from the German Scientific Council (DFG).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.M. (victor.malka@ensta.fr).

Reaction discovery enabled by DNA-templated synthesis and *in vitro* selection

Matthew W. Kanan, Mary M. Rozenman, Kaori Sakurai,
Thomas M. Snyder & David R. Liu

Department of Chemistry and Chemical Biology, Harvard University,
12 Oxford Street, Cambridge, Massachusetts 02138, USA

Current approaches to reaction discovery focus on one particular transformation. Typically, researchers choose substrates based on their predicted ability to serve as precursors for the target structure, then evaluate reaction conditions^{1–6} for their ability to effect product formation. This approach is ideal for addressing specific reactivity problems, but its focused nature might leave many areas of chemical reactivity unexplored. Here we report a reaction discovery approach that uses DNA-templated organic synthesis^{7–10} and *in vitro* selection to simultaneously evaluate many combinations of different substrates for bond-forming reactions in a single solution. Watson–Crick base pairing controls the effective molarities of substrates tethered to DNA strands; bond-forming substrate combinations are then revealed using *in vitro* selection for bond formation, PCR amplification and DNA microarray analysis. Using this approach, we discovered an efficient and mild carbon–carbon bond-forming reaction that generates an enone from an alkyne and alkene using an inorganic palladium catalyst. Although this approach is restricted to conditions and catalysts that are at least partially compatible with DNA, we expect that its versatility and efficiency will enable the discovery of additional reactions between a wide range of substrates.

A reaction discovery system capable of simultaneously evaluating many combinations of substrates for bond-forming reactivity in a single solution must meet several requirements. First, the system must organize complex substrate mixtures into discrete pairs that can react (or not react) without affecting the reactivity of the other substrate pairs. Second, the system requires a general method for separating reactive substrate pairs from unreactive pairs. Last, the reactive substrate pairs must be identified efficiently.

Recent developments in DNA-templated organic synthesis^{7–10} indicate that DNA annealing can organize many substrates in a single solution into DNA sequence-programmed pairs. To this end, we prepared two pools of DNA-linked substrates, with n substrates in pool A and m substrates in pool B. Each substrate in pool A is covalently linked to the 5' end of a set of DNA oligonucleotides containing one 'coding region' (uniquely identifying that substrate) and one of m different 'annealing regions' (Fig. 1a). Each of the m substrates in pool B is attached to the 3' end of an oligonucleotide containing a coding region that uniquely identifies the substrate and complements one of the m annealing regions in pool A.

When pools A and B are combined in a single aqueous solution at nanomolar concentrations, Watson–Crick base pairing organizes the mixture into $n \times m$ discrete pairs of substrates attached to complementary sequences. Only substrates linked to complementary oligonucleotides experience effective molarities in the millimolar range⁸; substrates linked to non-complementary oligonucleotides experience nanomolar solution concentrations and hence do not react with each other at a significant rate. This effective molarity-based design enables reactions that might otherwise be suppressed by the preferential dimerization of one or both substrates. The possibility of interference by the structure of DNA during reactions between substrates is minimized by using long and flexible substrate–DNA linkers¹¹.

The separation of reactive pairs of substrates is based on selection concepts used in the directed evolution of catalytic RNA and DNA^{12,13} (Fig. 1b). We covalently linked each pool B substrate to its corresponding oligonucleotide by a linker containing a biotin group and a disulphide bond (Fig. 1b and Supplementary Information). After incubation under a set of chosen reaction conditions, followed by cleavage of the disulphide bonds, only pool A sequences encoding bond formation between a pool A and pool B substrate remain covalently linked to biotin. Streptavidin affinity selection of the resulting solution separates biotinylated from non-biotinylated sequences. In contrast to many existing reaction discovery screens^{1,5} focused on a particular reaction type, our selection-based approach does not depend on any specific substrate or product property but instead can identify all substrate pairs capable of forming a covalent bond under the reaction conditions.

Reactive substrate pairs are identified by amplification of the selected sequences by polymerase chain reaction (PCR) followed by DNA microarray analysis (see below). Because PCR amplification is extremely sensitive, femtomole quantities of substrates are sufficient for the entire reaction discovery process.

We prepared pool A and pool B containing 12 substrates each (Fig. 2a), representing 144 heterocoupling combinations (see Supplementary Methods and Supplementary Fig. S1 for details). We also prepared DNA-linked substrates to enable the detection of homocoupling of any of the 24 different substrates, bringing the total number of unique substrate combinations to 168 (Fig. 2a). The 24 substrates in Fig. 2a were chosen to represent simple functional groups commonly encountered in organic molecules. New reactions forming bonds between simple functionalities are of special interest because they might provide more accessible alternatives to coupling reactions that require more complex substrates. Although our approach requires the preparation of DNA-linked substrates, a single nanomole-scale preparation of each pool provided sufficient material for more than 1,000 reaction discovery experiments that could collectively evaluate more than 168,000 combinations of substrates and reaction conditions.

To test the ability of our system to detect a single reactive combination of substrates out of 168 possibilities, we combined pool A and pool B in the presence of Cu(I), conditions known to promote a cycloaddition between a terminal alkyne (A5 in Fig. 2a) and an azide (B9 in Fig. 2a)¹⁴. Pool A members (2 pmol total) were combined with 2 pmol total of pool B members (12 fmol of material encoding each possible substrate combination) in the presence of 500 μ M Cu(I) in a total volume of 12 μ l. After 10 min at 25 °C, the salts were removed and disulphide linkages were cleaved with 0.1 M tris-carboxyethylphosphine hydrochloride. A control experiment lacking Cu(I) but identical in all other respects was also performed.

Sequences encoding bond-forming substrate pairs were captured with streptavidin-linked magnetic particles and amplified by PCR with a DNA primer labelled with the cyanine fluorophore Cy3. For comparison, an aliquot of the pool A sequences before selection was amplified by PCR with a Cy5-labelled primer. The Cy3-labelled and Cy5-labelled PCR products were combined and hybridized to a DNA microarray containing all 168 possible reaction-encoding sequences. The ratios of Cy3 (green) to Cy5 (red) fluorescence for all array locations were calculated and ordered by rank, and spots with green/red fluorescence ratios significantly higher than the majority of spots (in the experiments below, ratios above 1.5) were considered to be positive. A pre-quantified internal standard (bottom right corner of the array images in Fig. 2b) corresponding to a moderate level of reactivity was used as a positive control and as a reference for comparing different arrays.

For the experiment performed in the presence of Cu(I), the array spot corresponding to the combination of the alkyne (A5) and the azide (B9) was the sole spot that had a significant green/red fluorescence ratio (A5 + B9 = 8.5, standard = 2.8; see Fig. 2b and Supplementary Data). In contrast, the experiment performed

in the absence of any added metal yielded no green array spots other than the standard (Fig. 2b). These results show that the above method can detect a single bond-forming substrate combination from a complex mixture of 168 combinations, and confirm the high chemoselectivity characteristic of the Sharpless-modified Huisgen cycloaddition reaction¹⁵.

Cu(I) damages DNA by promoting radical-mediated processes¹⁶. Indeed, we observed 48% degradation of a DNA oligonucleotide exposed for 10 min to the Cu(I)-containing conditions (Supplementary Data). The successful detection of the alkyne–azide cycloaddition under conditions that degrade DNA shows that reaction conditions do not need to be fully DNA compatible, because the extreme sensitivity of PCR amplification¹⁷ enables bond formation to be revealed even if a significant fraction of the total DNA in an experiment is destroyed.

To validate this reaction discovery system further, we performed a selection for bond formation after exposure to organic reagents (1-(3-(dimethylamino)propyl)-3-ethylcarbodiimide hydrochloride and *N*-hydroxysuccinimide at pH 6.0) known to promote the coupling of amines with carboxylic acids¹⁸. Microarray analysis revealed one positive; this spot indeed corresponded to the combination of the carboxylic acid from pool A (A10) and the amine from pool B (B12) ($A_{10} + B_{12} = 6.9$, standard = 15.6; see Fig. 2b and Supplementary Data).

After successfully ‘rediscovering’ known bond-forming reactions, we examined the reactivity of a simple Pd(II) salt. Inorganic Pd salts and complexes containing Pd are known to mediate diverse coupling reactions¹⁹, indicating that Pd can activate a variety of organic

functional groups. Selection for bond formation in the presence of 500 μ M Na_2PdCl_4 for 1 h at 37 °C resulted in five strong positives ($A_7 + B_3$, $A_5 + B_3$, $A_4 + B_8$, $A_5 + B_5$ and A_5 homocoupling) with significant green/red fluorescence ratios (3.6 to 2.6; standard = 2.9). In addition, we observed five weaker positives ($A_9 + B_3$, $A_8 + B_3$, $A_8 + B_8$, $A_5 + B_8$ and $A_5 + B_9$) with lower, but possibly significant, green/red fluorescence ratios ranging from 1.9 to 1.6 (Fig. 2b and Supplementary Data). One positive representing the combination of aryl iodide (A_7) and acrylamide (B_3) is consistent with the well-known Heck reaction^{20,21} (assuming the formation of some Pd(0) under the reaction conditions). Other positives could be rationalized with mechanistic steps preceded in known Pd-mediated chemistries. For example, the spot indicating bond formation between an olefin (A_4) and an aryl boronic acid (B_8) is consistent with transmetalation of Pd(II) by the aryl boronic acid²² followed by the insertion of an olefin into the Pd–aryl bond and subsequent β -hydride elimination²³.

To determine whether these array results reflect genuine bond-forming events, we examined the putative reactions corresponding to the above ten spots in separate DNA-templated reactions. Denaturing polyacrylamide gel electrophoresis (PAGE) analysis indicated that all five strong positives and three of the five weak positives corresponded to authentic DNA-templated reactions, whereas two weak positives ($A_9 + B_3$ and $A_5 + B_9$) showed little or no product formation (Fig. 3 and Supplementary Fig. S2). We observed no product formation in control reactions in which Na_2PdCl_4 was omitted, or in which either of the reactive substrates was replaced with an unreactive alkane group. These findings are

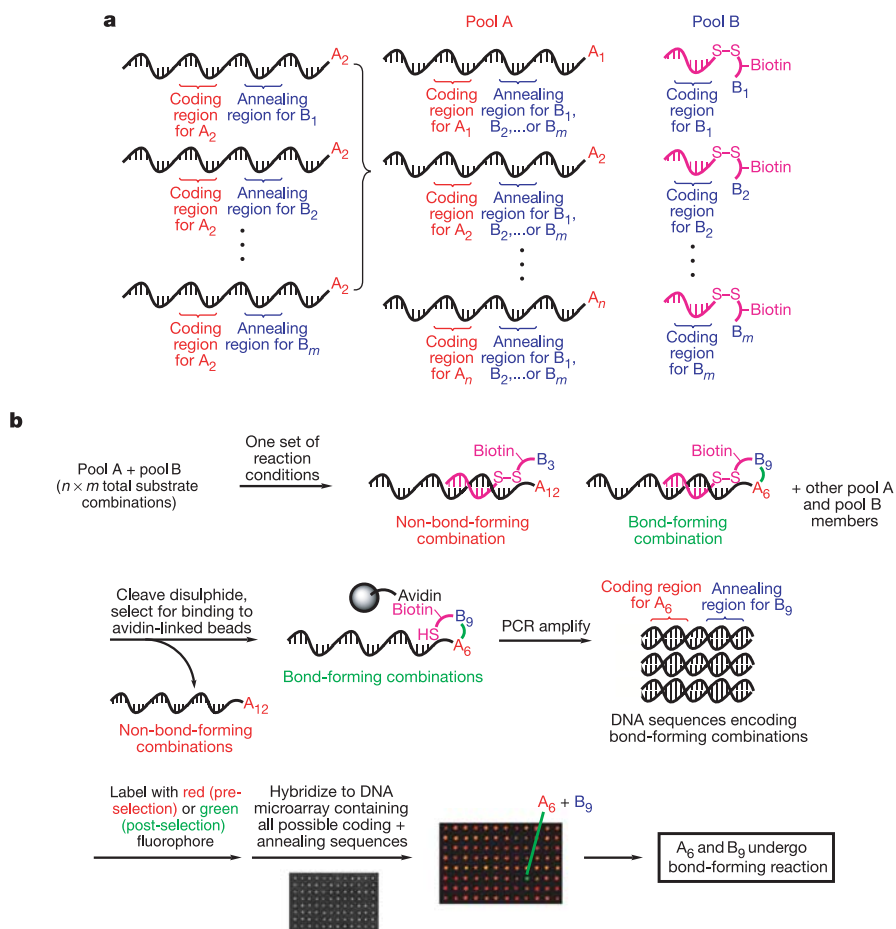


Figure 1 Key elements of a new approach to reaction discovery. **a**, Two pools of DNA-linked organic functional groups that associate each of $n \times m$ substrate

combinations with a unique DNA sequence. **b**, A general one-pot selection and analysis method for the detection of bond-forming reactions between DNA-linked substrates.

consistent with the detection of Pd-dependent reactions that couple substrate groups and do not couple the functionality present in DNA.

We further characterized the PAGE-validated Pd-mediated DNA-templated reactions using matrix assisted laser desorption/ionization–time-of-flight (MALDI–TOF) mass spectrometry (Fig. 3). The observed product masses are consistent with covalent bond formation. We suggest possible structures consistent with the observed masses for each of the reaction products but note that other products or mixtures of products are also consistent with the data. Together with the PAGE analysis, these results show that this reaction discovery system reliably and efficiently reveals bond-forming events even under conditions that cause multiple combinations of substrates to react.

Increasing the stringency of the reaction conditions by decreasing the temperature to 25 °C and quenching the reaction after 20 min

decreased the number of strong positives to four (A5+B5, A5+B3, A4+B8 and A5 homocoupling; green/red fluorescence ratios = 3.7 to 2.7; standard = 3.5) and the number of weak positives to two (A5+B8 and A5+B9; fluorescence ratios = 1.9 and 1.7; see Fig. 2b and Supplementary Data). Consistent with these results, the five authentic bond-forming reactions among this set (all four strong positives and weak positive A5+B8) showed substantial product formation by PAGE analysis when performed under the 25 °C conditions, whereas the three additional reactions listed in Fig. 3 but not appearing as positives in the 25 °C experiment (A8+B8, A8+B3 and A7+B3) showed significantly weaker product formation (Fig. 3). Varying the stringency of the reaction conditions before selection can therefore efficiently distinguish substrate combinations on the basis of their level of reactivity.

The above results suggested that the Pd(II)-mediated carbon–carbon bond formation between a simple terminal alkyne and

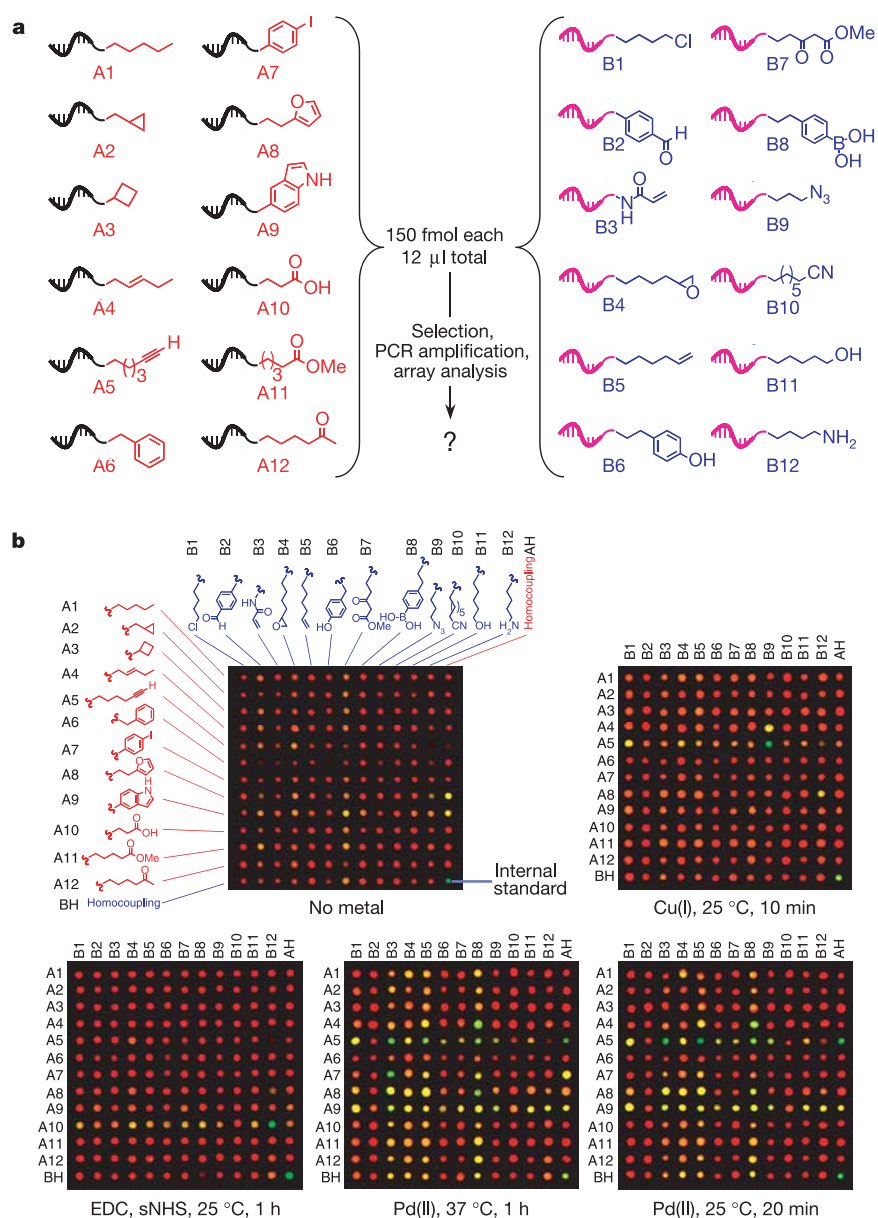


Figure 2 Results from reaction discovery selections and analysis. **a**, Pool A and pool B substrates used in this work. **b**, Qualitative results of reaction discovery selections after exposure to the reaction conditions listed below each array image. Spots that are significantly green suggest bond formation between the corresponding substrates

(quantitative fluorescence ratios in Fig. 3 and the Supplementary Data are used for actual interpretations). The 840 reaction possibilities in these five experiments were evaluated by one researcher in two days. See Supplementary Methods for detailed reaction conditions.

Substrates		Green/red fluorescence ratios		DNA-templated yields (%)		Product consistent with observed mass
		37°C	25°C	37°C	25°C	
A5	B5	2.7	3.7	35	31	
A5	B3	3.5	3.1	28	20	
A5	B8	1.6	1.9	36	34	
A5	Homocoupling	2.6	2.7	45	42	
A4	B8	3.0	2.8	57	39	
A8	B8	1.8	<1.2	30	10	
A8	B3	1.8	<1.2	19	<10	
A7	B3	3.6	<1.2	39	14	

Figure 3 Characterization in a DNA-templated format of array positives resulting from exposure to 500 μ M Na_2PdCl_4 at 37 °C for 1 h, or at 25 °C for 20 min. Putative reactions were screened by PAGE and MALDI-TOF mass spectrometric analysis (Supplementary Information). Template architectures and reaction conditions were chosen to match those

used in the selection rather than to maximize product yields. Product structures other than those proposed are possible. The green/red fluorescence ratio for the internal standard at 37 °C and 25 °C is 2.9 and 3.5, respectively. For A8 + B3, reactivity was not sufficient to obtain reliable product mass characterization.

Entry	Metal(s)	Solvent	Conditions	Isolated yield
a	1 equiv. Na_2PdCl_4	1 M NaCl in H_2O	25 °C, 15 h	86%
b	5 mol% Na_2PdCl_4 1 equiv. CuCl_2	100 mM NaCl in H_2O	25 °C, 2 h	90%
c	5 mol% Na_2PdCl_4 1 equiv. CuCl_2	9:1 THF: H_2O	25 °C, 4 h	91%
d	15 mol% Na_2PdCl_4 1 atm O_2	9:1 THF: H_2O	25 °C, 14 h	73%
e	1 equiv. CuCl_2	100 mM NaCl in H_2O	25 °C, 4 h	0%
f	1 equiv. CuCl	100 mM NaCl in H_2O	25 °C, 4 h	0%

Figure 4 Characterization of a new alkyne-alkene macrocyclization reaction in a non-DNA-templated format. The macrocyclic enone product (**2**) was characterized by ^1H -NMR, ^{13}C -NMR, COSY, UV-visible spectrometry and high-resolution mass electrospray (Supplementary Information). We speculate that product formation proceeds through the following sequence: soft deprotonation of the alkyne to form a $\text{Pd}(\text{II})$ -alkynyl intermediate; insertion of the alkene into the Pd -alkyne bond; β -hydride elimination to form a conjugated enyne; $\text{Pd}(\text{II})$ -catalysed hydration of the alkyne to form an enol π -allyl Pd complex; enol tautomerization and π -allyl Pd protonation to generate the *trans*-enone.

terminal alkene (A5+B5) proceeds efficiently to generate a possible enone product (Fig. 2b and Fig. 3), prompting a detailed investigation of this reaction. We synthesized a small-molecule substrate (**1**) to study a non-DNA-templated, intramolecular version of this alkyne-alkene coupling reaction on a multi-milligram scale (Fig. 4). The addition of **1** to one equivalent of $\text{Pd}(\text{II})$ in 1 M aqueous NaCl over 15 h followed by purification by reverse-phase high-

performance liquid chromatography provided the 20-membered macrocyclic *trans*-enone **2** as a single olefin stereoisomer in 86% isolated yield (Fig. 4, entry a). The structure of **2** was confirmed by ^1H -NMR, ^{13}C -NMR, homonuclear correlated spectroscopy (COSY) and high-resolution mass spectrometry (Supplementary Data). To our knowledge, this is the first example of macrocyclic enone formation from a simple alkyne-alkene precursor. The finding also shows that a reaction discovered by a selection for DNA-templated covalent bond formation can operate in a non-DNA-templated format on a much larger (10^9 -fold) scale.

Because the formation of an enone from an alkyne and an alkene represents an oxidative coupling, we proposed that the reaction could be performed with catalytic quantities of Pd by introducing an oxidant such as CuCl_2 + air to reoxidize $\text{Pd}(0)$ to $\text{Pd}(\text{II})$ (ref. 24). Indeed, the addition of **1** to 5 mol% Pd and 1 equivalent of CuCl_2 over 2 h in water at 25 °C provided enone **2** as the sole observed product in 90% isolated yield (Fig. 4, entry b). Significantly, this reaction maintains its high efficiency in solvent containing 9:1 tetrahydrofuran:water (Fig. 4, entry c). Product was formed in similar yields, although at a slower rate, using 1 atm O_2 instead of CuCl_2 to reoxidize $\text{Pd}(0)$ (Fig. 4, entry d). Control reactions with CuCl_2 or CuCl alone yielded no observed product formation (Fig. 4, entries e and f).

The discovery of this alkyne-alkene coupling reaction suggests the value of searching a large number of substrate combinations for unexpected reactions. While aqueous $\text{Pd}(\text{II})$ has been known for more than 40 years to oxidize alkenes rapidly to ketones^{24,25}, which are unreactive towards alkynes under these conditions (Fig. 2b and Supplementary Data below Supplementary Fig. S4), our approach to reaction discovery revealed that carbon-carbon bond formation between alkynes and alkenes can outcompete alkene oxidation. Although other enone-forming coupling reactions such as the Horner-Wadsworth-Emmons reaction²⁶ or aldol condensation are known, the mild reaction conditions, simple hydrocarbon starting materials and high efficiency of the transformation discovered here might render it an attractive alternative for addressing

some macrocyclization problems in organic synthesis. In addition, the compatibility of this reaction with both organic and aqueous solvents might facilitate the synthesis of highly functionalized macrocyclic enones, of which there are numerous known biologically active examples²⁷.

Six of the eight observed heterocoupling reactions in Fig. 3 involve a substrate that couples with itself under the reaction conditions (reactions involving either the alkyne or the aryl boronic acid; see Supplementary Data below Supplementary Fig. S4). In a DNA-templated format, heterocoupling without competitive homocoupling is possible because a substrate experiences its sequence-programmed heterocoupling partner at a much higher effective concentration than the concentration of another identical substrate molecule.

Biological macromolecules have previously been used to address specific problems in chemical reactivity^{28–30}. In contrast, our approach uses the ability of nucleic acids to direct effective molarities and undergo *in vitro* selection and amplification to reveal bond-forming reactivity in a general manner. Once DNA-linked substrate pools are prepared, a single researcher can evaluate thousands of combinations of substrates and reaction conditions in a two-day experiment. Our approach requires water-soluble catalysts and aqueous reaction conditions that are at least partially compatible with DNA; however, as illustrated by the alkyne–alkene coupling reaction, some of the discovered reactions will proceed in a non-DNA-templated format in organic solvents. We expect that using this approach in a broad examination of reaction conditions including transition-metal complexes, Lewis acids, mild oxidants or reductants, and organic reagents will lead to the discovery of additional bond-forming reactions between simple and relatively unreactive functional groups. □

Received 11 June; accepted 3 August 2004; doi:10.1038/nature02920.

1. Stambuli, J. P. & Hartwig, J. F. Recent advances in the discovery of organometallic catalysts using high-throughput screening assays. *Curr. Opin. Chem. Biol.* **7**, 420–426 (2003).
2. Reetz, M. T. Combinatorial and evolution-based methods in the creation of enantioselective catalysts. *Angew. Chem. Int. Edn Engl.* **40**, 284–310 (2001).
3. Stambuli, J. P., Stauffer, S. R., Shaughnessy, K. H. & Hartwig, J. F. Screening of homogeneous catalysts by fluorescence resonance energy transfer. Identification of catalysts for room-temperature Heck reactions. *J. Am. Chem. Soc.* **123**, 2677–2678 (2001).
4. Taylor, S. J. & Morken, J. P. Thermographic selection of effective catalysts from an encoded polymer-bound library. *Science* **280**, 267–270 (1998).
5. Lober, O., Kawatsura, M. & Hartwig, J. F. Palladium-catalyzed hydroamination of 1,3-dienes: a colorimetric assay and enantioselective additions. *J. Am. Chem. Soc.* **123**, 4366–4367 (2001).
6. Evans, C. A. & Miller, S. J. Proton-activated fluorescence as a tool for simultaneous screening of combinatorial chemical reactions. *Curr. Opin. Chem. Biol.* **6**, 333–338 (2002).
7. Calderone, C. T., Puckett, J. W., Gartner, Z. J. & Liu, D. R. Directing otherwise incompatible reactions in a single solution by using DNA-templated organic synthesis. *Angew. Chem. Int. Edn Engl.* **41**, 4104–4108 (2002).
8. Gartner, Z. J. & Liu, D. R. The generality of DNA-templated synthesis as a basis for evolving non-natural small molecules. *J. Am. Chem. Soc.* **123**, 6961–6963 (2001).
9. Gartner, Z. J., Kanan, M. W. & Liu, D. R. Expanding the reaction scope of DNA-templated synthesis. *Angew. Chem. Int. Edn Engl.* **41**, 1796–1800 (2002).
10. Rosenbaum, D. M. & Liu, D. R. Efficient and sequence-specific DNA-templated polymerization of peptide nucleic acid aldehydes. *J. Am. Chem. Soc.* **125**, 13924–13925 (2003).
11. Li, X. & Liu, D. R. Stereoselectivity in DNA-templated organic synthesis and its origins. *J. Am. Chem. Soc.* **125**, 10188–10189 (2003).
12. Wilson, D. S. & Szostak, J. W. *In vitro* selection of functional nucleic acids. *Annu. Rev. Biochem.* **68**, 611–647 (1999).
13. Joyce, G. F. Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem.* **73**, 791–836 (2004).
14. Rostovtsev, V. V., Green, L. G., Fokin, V. V. & Sharpless, K. B. A stepwise Huisgen cycloaddition process: copper(I)-catalyzed regioselective 'ligation' of azides and terminal alkynes. *Angew. Chem. Int. Edn Engl.* **41**, 2596–2599 (2002).
15. Wang, Q. *et al.* Bioconjugation by copper(I)-catalyzed azide–alkyne [3+2] cycloaddition. *J. Am. Chem. Soc.* **125**, 3192–3193 (2003).
16. Burrows, C. J. & Müller, J. G. Oxidative nucleobase modifications leading to strand scission. *Chem. Rev.* **98**, 1109–1152 (1998).
17. Kramer, M. F. & Coen, D. M. in *Current Protocols in Molecular Biology* Vol. 2 (eds Ausubel, F. M. *et al.*) 15.1.1–15.1.3 (Wiley, New York, 2001).
18. Bailey, P. D., Collier, I. D. & Morgan, K. M. in *Comprehensive Organic Functional Group Transformations* Vol. 5 (eds Katritzky, A. R., Meth-Cohn, O. & Rees, C. W.) 257–307 (Pergamon, New York, 1995).
19. Tsuji, J. *Palladium Reagents and Catalysts* (Wiley, New York, 1995).
20. Heck, R. F. Palladium-catalyzed vinylation of organic halides. *Org. React.* **27**, 345–390 (1982).
21. Li, C.-J. & Chan, T.-H. *Organic Reactions in Aqueous Media* (Wiley, New York, 1997).
22. Miyauchi, N. & Suzuki, A. Palladium-catalyzed cross-coupling reactions of organoboron compounds. *Chem. Rev.* **95**, 2457–2483 (1995).

23. Crabtree, R. H. *The Organometallic Chemistry of the Transition Metals* (Wiley, New York, 2001).
24. Smidt, J. *et al.* Olefinoxydation mit palladiumchlorid-katalysatoren. *Angew. Chem.* **74**, 93–102 (1962).
25. Smidt, J. *et al.* Katalytische umsetzungen von olefinen an platinmetall-verbindungen: das consortium-verfahren zur herstellung von acetaldehyd. *Angew. Chem.* **71**, 176–182 (1959).
26. Wadsworth, W. S. J. & Emmons, W. D. Utility of phosphonate carbanions in olefin synthesis. *J. Am. Chem. Soc.* **83**, 1733–1738 (1961).
27. Staunton, J. & Weissman, K. J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **18**, 380–416 (2001).
28. Kohli, R. M., Walsh, C. T. & Burkart, M. D. Biomimetic synthesis and optimization of cyclic peptide antibiotics. *Nature* **418**, 658–661 (2002).
29. Breslow, R. Biomimetic chemistry and artificial enzymes: catalysis by design. *Acc. Chem. Res.* **28**, 146–153 (1995).
30. Schultz, P. G. & Lerner, R. A. Completing the circle. *Nature* **418**, 485 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Kehayova for the initial analysis of early reaction discovery selections, and C. Bailey (Bauer Center for Genomics Research) for assistance with DNA microarray preparation and analysis. We thank DNA Software for assistance with screening coding and annealing sequences. This work was supported by the National Institutes of Health, the Office of Naval Research, and the Arnold and Mabel Beckman Foundation. M.W.K. and T.M.S. are NSF Graduate Research Fellows. M.M.R. is an NDSEG Graduate Research Fellow. K.S. is a Helen Hay Whitney Postdoctoral Research Fellow.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to D.R.L. (liu@chemistry.harvard.edu).

Photosynthetic microbial mats in the 3,416-Myr-old ocean

Michael M. Tice & Donald R. Lowe

Department of Geological and Environmental Sciences, Stanford University, Stanford, California 94305, USA

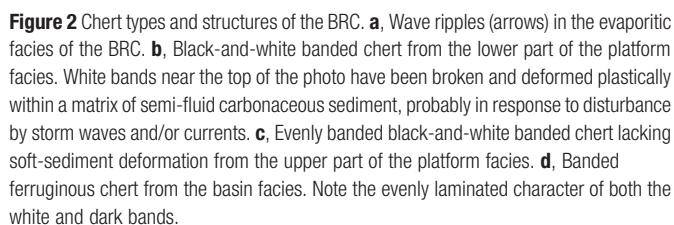
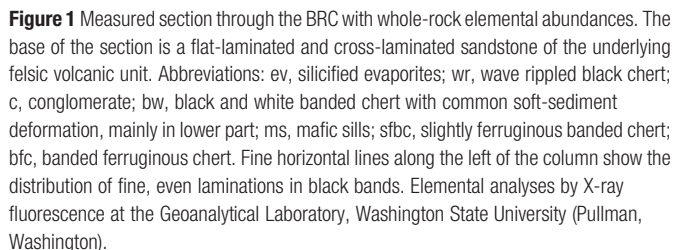
Recent re-evaluations of the geological record of the earliest life on Earth have led to the suggestion that some of the oldest putative microfossils¹ and carbonaceous matter were formed through abiotic hydrothermal processes^{2,3}. Similarly, many early Archaean (more than 3,400-Myr-old) cherts have been reinterpreted as hydrothermal deposits rather than products of normal marine sedimentary processes^{2,4,5}. Here we present the results of a field, petrographic and geochemical study testing these hypotheses for the 3,416-Myr-old Buck Reef Chert, South Africa. From sedimentary structures and distributions of sand and mud, we infer that deposition occurred in normal open shallow to deep marine environments. The siderite enrichment that we observe in deep-water sediments is consistent with a stratified early ocean^{6,7}. We show that most carbonaceous matter was formed by photosynthetic mats within the euphotic zone and distributed as detrital matter by waves and currents to surrounding environments. We find no evidence that hydrothermal processes had any direct role in the deposition of either the carbonaceous matter or the enclosing sediments. Instead, we conclude that photosynthetic organisms had evolved and were living in a stratified ocean supersaturated in dissolved silica^{8,9} 3,416 Myr ago.

The Buck Reef Chert (BRC) is a 250–400-m-thick unit of carbonaceous and ferruginous chert that regionally marks the base of the Kromberg Formation in the Barberton greenstone belt, South Africa¹⁰ (Fig. 1). Along the west limb of the Onverwacht anticline, it includes three main facies reflecting deposition under progressively deeper-water conditions: first, a basal evaporite facies 5–40 m thick that interfingers with the underlying felsic sandstone

The evaporite facies consists of silicified laminated and wave-rippled sediment (Fig. 2a), silicified evaporitic layers originally composed of nahcolite (NaHCO_3) and solution-collapse layer after evaporites¹¹. Laminated and wave-rippled units are composed mainly of microquartz containing detrital carbonaceous particles and altered felsic volcanic grains. Aluminium, zirconium and chromium abundances are moderate (Fig. 1). These elements are relatively immobile during low-grade metamorphism, making them useful tracers of clastic inputs in similar rocks⁸. Their abundances and relative ratios in evaporite facies rocks reflect clastic input from surrounding volcanic units. The underlying sandstone has been interpreted as a coastal and braidplan deposit¹¹. Evaporites were deposited in shallow marginal-marine ponds subject to gentle surface waves¹¹. Carbonaceous matter occurs as detrital grains of silt

The contact between the evaporite and overlying platform facies is marked by a thin conglomerate, probably a transgressive lag left during regional marine flooding. Much of the shallow-water coastal environment is not represented by deposits because available sediments were easily eroded within these high-energy settings. The platform facies crops out for more than 50 km along strike and is composed largely of black-and-white banded chert made up of bands of black carbonaceous chert less than 1 cm to about 15 cm thick alternating with bands of pure, white-weathering, translucent chert from 1 mm to 10 cm thick (Fig. 2b, c). In the lower half of the platform facies, black bands are composed largely of thin current-deposited layers of rounded, detrital silica and carbonaceous grains of fine sand size (less than 0.25 mm) to granule size (4–5 mm), including irregular ripped up chunks of carbonaceous matter (Fig. 3a), and minor siderite (FeCO_3).

The abundance of rounded, sand-sized, detrital carbonaceous particles suggests that debris in the lower platform section was



moved and sorted under the influence of weak waves or currents. The unit's wide extent, abundance of low-density detrital grains and absence of coarse clastic particles and high-energy current structures indicate deposition on an open marine wave-active and current-active shelf. Most banded sediment was disrupted by early soft-sediment flowage and deformation to form breccias originally composed of rigid plates to irregular soft plastically deformed masses of white chert within a fluid matrix of black chert (Fig. 2b). This soft-sediment disruption is interpreted to reflect the effects of storm events, which set up internal stresses and mixing within the still soft, gelatinous bottom materials rich in silica and organic compounds.

In the upper half of the platform facies, soft sediment deformation and brecciation are less common (Fig. 2c). Carbonaceous bands are mainly finely and evenly flat laminated and lack well-developed coarse detrital textures. In thin section they are composed mainly of fine, irregular, flattened carbonaceous grains (Fig. 3b) and flattened silica particles. Fine mat-like laminations and microfossils are absent. Bands show little or no evidence of current or wave activity or sorting of carbonaceous grains. Waves, currents and larger-scale storm activity that affected the bottom were infrequent.

Aluminium, zirconium and chromium abundances in platform facies cherts are low relative to those in evaporite facies cherts. This decrease reflects the disappearance of clastic debris as the volcanic edifice subsided below sea level. Iron abundances are also low (0.02–0.3 wt% FeO).

The overlying basin facies is composed of banded slightly ferruginous chert to banded ferruginous chert consisting of alternating bands of relatively pure white-weathering chert, 1 mm to 2 cm thick, and dark rust-coloured sediment, less than 2 cm thick (Fig. 2d). Detrital particles, soft sediment deformation and brecciation are rare to absent. Dark ferruginous bands are finely laminated and at the surface are composed largely of intergrown microquartz, hydrous ferric oxide and goethite, although siderite is present in relatively unweathered enclaves. In subsurface samples of similar rocks, the principal iron-bearing mineral is siderite; iron oxides are

absent. The main iron-bearing mineral in the basin facies was probably siderite that has been oxidized near the present surface by modern groundwater¹⁶. Siderite is present as silt-sized equant grains and fine rhombic grains. They are disseminated, define laminations, form lenses and are found both isolated and associated with carbonaceous matter. The occurrence of siderite isolated from carbonaceous matter suggests that it was not formed by the reduction of early iron oxyhydroxides. Carbonaceous content is low in surface samples, probably reflecting removal by modern oxidation, but subsurface samples contain abundant fine particulate carbonaceous matter. Aluminium, zirconium and chromium abundances are variable but generally higher than in the platform facies (Fig. 1), probably reflecting enrichment of these sediments in windblown ash or dust because of overall low sedimentation rates. Iron abundances are two orders of magnitude higher than in platform facies cherts (1–50 wt%). The basin facies was deposited in an extremely low-energy environment, far below storm wave base. It represents the slow accumulation by settling of siderite, silica, fine airborne clastic material and fine hemipelagic carbonaceous debris.

Whereas the presence of nahcolite as a primary evaporite testifies to high bicarbonate levels in shallow water¹¹, the lack of abundant iron precipitates in evaporite and platform facies sediments indicates that iron concentrations in shallow water were low. The abundance of siderite in basin facies cherts implies siderite saturation in deep water. We suggest that the abundance of siderite in deep-water deposits and its paucity or absence in shallow-water deposits reflect stratification of the Archaean ocean into a shallow, CO₂-dominated layer and a deep, iron-rich layer. Siderite precipitated at the mixing interface between these masses and accumulated in deep basin sediments. This inference is consistent with previous models of Archaean ocean stratification based on studies in other areas^{6,7}.

Although some investigators have interpreted the BRC and similar chert units as hydrothermal exhalites, hydrothermal injections or hydrothermally modified sediments^{2,4,5,17}, the BRC shows clear evidence for deposition as normal marine sediments and silicification through interaction with marine waters supersaturated with amorphous silica⁸. The unit ranges from 250 to 400 m thick and crops out for more than 50 km along strike¹¹. It exhibits facies deposited in coastal, evaporitic, shelf and deep-basin settings. Such persistence in space and development across a range of depositional

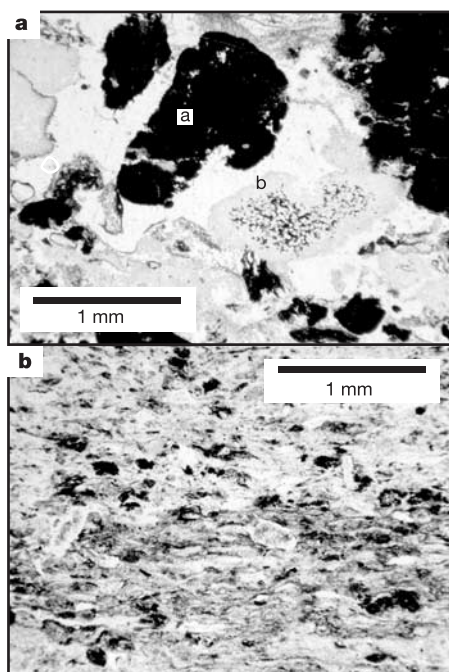


Figure 3 Grains and textures of the BRC. **a**, Coarse carbonaceous (a) and silica (b) grains from the lower part of the platform facies. **b**, Fine compacted carbonaceous and silica grains from the upper part of platform facies.

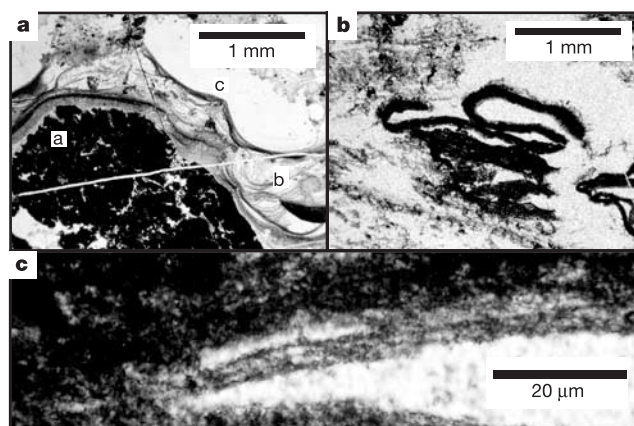


Figure 4 Laminated carbonaceous matter and filaments in the lower part of the platform facies. **a**, Dark carbonaceous laminations draping an underlying coarse detrital carbonaceous grain (a), showing internal anastomosing and draping character (b) and, at the top (c) draping irregularities in underlying carbonaceous laminations. **b**, Dark carbonaceous laminations that have been eroded and rolled up by currents. **c**, Bundled filaments in the rolled laminations in **b**.

environments is unlikely for a hydrothermal system. There are no facies or thickness trends that indicate the existence of local sources of silica or other materials, such as hydrothermal vents, and no cross-cutting features or deposits that might represent syndepositional hydrothermal conduits or hot springs. Instead, silicification most probably resulted from precipitation out of a hot ocean¹⁸ that had no biological sink for silica⁹.

Mat-like laminations are restricted to shallow-water environments in an apparent ecological control of their distribution. Carbonaceous matter in the BRC has a carbon isotopic composition of -35 to -20% compared with PDB¹⁴. Although mechanisms have been proposed by which hydrothermal systems could produce isotopically fractionated methane¹⁹, no hydrothermal systems capable of generating large volumes of fractionated carbonaceous matter consistently within this isotopic range are known.

The isotopic composition of BRC carbonaceous matter is consistent with fixation by autotrophs employing the Calvin cycle²⁰. Organisms with a variety of physiologies use this pathway, including some types of oxygenic and anoxygenic photosynthesizers, and many chemoautotrophs such as sulphide, iron and hydrogen oxidizers²¹. The absence of ferric oxides in the platform facies implies that carbon was not fixed predominantly by iron oxidation. Sulphide and hydrogen oxidation both require free O₂. The presence of siderite and the absence of ferric oxides throughout the BRC indicates that the partial pressure of O₂ was very low, making both of these metabolisms unlikely as primary carbon-fixation pathways. During deposition of the BRC, mat-forming microorganisms were apparently confined to shallow-water settings above deep storm wave base. In modern oceans, this corresponds to depths of less than 200 m (ref. 22). Restriction of these communities to shallow water probably reflects confinement to the euphotic zone, which generally corresponds to depths of less than 150 m (ref. 23). Taken together, the carbon isotopic composition of BRC carbonaceous matter, the presence of siderite and lack of primary ferric oxides, and the restriction of microbial mats to shallow water indicate that photosynthetic, probably anoxygenic, microbes were active in the 3,416-Myr-old ocean.

Detrital carbonaceous grains become simpler in morphology and finer in size with the transition from shallow platform settings to deep platform and basin settings. This distribution is strongly indicative of a shallow-water origin for these grains, most probably ultimately by the erosion of shallow-water mats. Mat material was ripped up and redistributed as detritus in shallow marine environments by waves and storms. In deeper waters, fine detrital carbonaceous matter accumulated through hemipelagic settling along with precipitated siderite, fine clastic material, and silica. The BRC provides a clear window into ocean chemistry and structure and biological processes active about 3.4 Gyr ago. □

Received 9 January; accepted 26 July 2004; doi:10.1038/nature02888.

- Schopf, J. W. & Packer, B. M. Early Archean (3.3 billion to 3.5 billion-year-old) microfossils from Warrawoona Group, Australia. *Science* **237**, 70–73 (1987).
- Brasier, M. D. *et al.* Questioning the evidence for Earth's oldest fossils. *Nature* **416**, 76–81 (2002).
- García-Ruiz, J. M. *et al.* Self-assembled silica-carbonate structures and detection of ancient microfossils. *Science* **302**, 1194–1197 (2003).
- Westall, F. *et al.* Early Archean fossil bacteria and biofilms in hydrothermally-influenced sediments from the Barberton greenstone belt, South Africa. *Precamb. Res.* **106**, 93–116 (2001).
- Paris, I., Stanistreet, I. G. & Hughes, M. J. Cherts of the Barberton greenstone belt as products of submarine exhalative activity. *J. Geol.* **93**, 111–129 (1985).
- Klein, C. & Beukes, N. J. Geochemistry and sedimentology of a facies transition from limestone to iron-formation deposition in the early Proterozoic Transvaal Supergroup, South Africa. *Econ. Geol.* **84**, 1733–1774 (1989).
- Lowe, D. R. in *Early Life on Earth. Nobel Symposium No. 84* (ed. Bengtson, S.) 24–35 (Columbia Univ. Press, New York, 1994).
- Lowe, D. R. in *Geologic Evolution of the Barberton Greenstone Belt, South Africa* (eds Lowe, D. R. & Byerly, G. R.) 83–114 (Geol. Soc. Am. Spec. Pap. 329, Boulder, Colorado, 1999).
- Siever, R. The silica cycle in the Precambrian. *Geochim. Cosmochim. Acta* **56**, 3265–3272 (1992).
- Lowe, D. R. & Byerly, G. R. in *Geologic Evolution of the Barberton Greenstone Belt, South Africa* (eds Lowe, D. R. & Byerly, G. R.) 1–36 (Geol. Soc. Am. Spec. Pap. 329, Boulder, Colorado, 1999).
- Lowe, D. R. & Fisher Worrell, G. in *Geologic Evolution of the Barberton Greenstone Belt, South Africa* (eds Lowe, D. R. & Byerly, G. R.) 167–188 (Geol. Soc. Am. Spec. Pap. 329, Boulder, Colorado, 1999).

- Tice, M. M., Bostick, B. C. & Lowe, D. R. Thermal history of the 3.5–3.2 Ga Onverwacht and Fig Tree Groups, Barberton greenstone belt, South Africa. *Geology* **32**, 37–40 (2004).
- Sumner, D. Y. Late Archean calcite-microbe interactions: Two morphologically distinct microbial communities that affected calcite nucleation differently. *Palaios* **12**, 302–318 (1997).
- Walsh, M. M. & Lowe, D. R. in *Geologic Evolution of the Barberton Greenstone Belt, South Africa* (eds Lowe, D. R. & Byerly, G. R.) 115–132 (Geol. Soc. Am. Spec. Pap. 329, Boulder, Colorado, 1999).
- Simonson, B. M., Schubel, K. A. & Hassler, S. W. Carbonate sedimentology of the early Precambrian Hamersley Group of Western Australia. *Precamb. Res.* **60**, 287–335 (1993).
- Lowe, D. R. & Byerly, G. R. Ironstone pods in the Archean Barberton greenstone belt, South Africa: Earth's oldest seafloor hydrothermal vents reinterpreted as Quaternary subaerial springs. *Geology* **31**, 909–912 (2003).
- de Wit, M. J., Hart, R., Martin, A. & Abbott, P. Archean abiogenic and probable biogenic structures associated with mineralized hydrothermal vent systems and regional metasomatism, with implications for greenstone belt studies. *Econ. Geol.* **77**, 1783–1802 (1982).
- Knauth, L. P. & Lowe, D. R. High Archean climatic temperature inferred from oxygen isotope geochemistry of cherts in the 3.5 Ga Swaziland Supergroup, South Africa. *Geol. Soc. Am. Bull.* **115**, 566–580 (2003).
- Horita, J. & Berndt, M. E. Abiogenic methane formation and isotopic fractionation under hydrothermal conditions. *Science* **285**, 1055–1057 (1999).
- Schidlowski, M. in *Microbial Sediments* (eds Riding, R. E. & Awramik, S. M.) 84–95 (Springer, New York, 2000).
- Madigan, M. T., Martinko, J. M. & Parker, J. *Brock Biology of Microorganisms* (Prentice-Hall, Upper Saddle River, New Jersey, 1997).
- Boggs, S. *Principles of Sedimentology and Stratigraphy* (Prentice-Hall, Englewood Cliffs, New Jersey, 1995).
- Lalli, C. M. & Parsons, T. R. *Biological Oceanography: An Introduction* (Butterworth-Heinemann, Boston, 1997).

Acknowledgements This work was supported by NASA Exobiology Program grants to D.L. M.T. was also supported by a William R. and Sara Hart Kimball Stanford Graduate Fellowship and by a Harvey Fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.T. (mtice@pangea.stanford.edu).

Excitation of Earth's continuous free oscillations by atmosphere-ocean-seafloor coupling

Junkee Rhie & Barbara Romanowicz

Berkeley Seismological Laboratory and Department of Earth and Planetary Science, University of California, Berkeley, California 94720, USA

The Earth undergoes continuous oscillations, and free oscillation peaks have been consistently identified in seismic records in the frequency range 2–7 mHz (refs 1, 2), on days without significant earthquakes. The level of daily excitation of this 'hum' is equivalent to that of magnitude 5.75 to 6.0 earthquakes^{3,4}, which cannot be explained by summing the contributions of small earthquakes^{1,3}. As slow or silent earthquakes have been ruled out as a source for the hum⁴ (except in a few isolated cases⁵), turbulent motions in the atmosphere or processes in the oceans have been invoked^{3,6–8} as the excitation mechanism. We have developed an array-based method to detect and locate sources of the excitation of the hum. Our results demonstrate that the Earth's hum originates mainly in the northern Pacific Ocean during Northern Hemisphere winter, and in the Southern oceans during Southern Hemisphere winter. We conclude that the Earth's hum is generated by the interaction between atmosphere, ocean and sea floor, probably through the conversion of storm energy to oceanic infragravity waves that interact with seafloor topography.

Elucidating the physical mechanism responsible for the continuous oscillations represents an intriguing scientific challenge. The source(s) should be close to the Earth's surface, as the fundamental

mode appears to be preferentially excited. One proposed mechanism relates the observed oscillations to random excitation by turbulent motions in the atmosphere^{3,6,7}. Amplitude levels and frequency dependence estimated stochastically, and constrained by actual barometer readings, are in agreement with the observed continuous oscillation levels. Also in support of this interpretation, seasonal variations in the level of energy present in the continuous oscillations have a six-month periodicity, with maxima in January and July corresponding to winter in Northern and Southern hemispheres respectively⁴, correlated with maxima in average atmospheric pressure variations. On the other hand, Nishida and Kobayashi⁹ proposed that the source might be distributed over the entire surface of the Earth. The fact that the background mode signal can be brought out even more clearly by correcting the observed spectra for signal correlated with local barograph recordings provides further evidence for the non-local character of the excitation process¹⁰. An alternative potential source of excitation of the 'hum' could be in the oceans, resulting from interaction between wind and ocean waves. Such an interpretation is supported by the similarity of the shape of ocean-bottom pressure spectra and ground-motion noise spectra⁸.

In order to make further progress on this issue, it is important to determine whether the excitation source is indeed distributed over

most of the Earth, or whether most of it occurs either in the oceans or on land. The approaches used so far, based on the computation of spectra, or the correlation of signals across full great-circle paths⁴, allow the detection but not the location of the sources. The latter must be addressed using a propagating wave methodology.

We have developed an array-based method to detect and locate sources of very-long-period surface wave energy, using the dispersive properties of Rayleigh waves¹¹. The use of surface waves to detect and locate earthquakes was proposed many decades ago¹². Generally, moderate size events and relatively short periods (that is, 20–100 s) have been considered¹³. Array methods have been developed and applied widely for detection and analysis of body waves in the azimuth–slowness domain¹⁴, and for the analysis of sources of microseisms^{15,16}. Recently, Ekström *et al.*¹⁷ developed a stacking technique based on a global network of ~100 long-period seismic stations and, in the period band 35–150 s, they detected many glacial earthquakes depleted in high-frequency energy. On the other hand, Nishida *et al.*¹⁸ showed that the vertical seismic background noise in the entire pass-band from 2 to 20 mHz is dominated by globally propagating Rayleigh waves. Here we utilize the dispersive properties of mantle Rayleigh waves in the period band 150–500 s across two regional networks of very broadband seismometers, one in Japan (F-net) and the other in California (BDSN). Our analysis is centred around a period of 240 s, where the dispersion of Rayleigh waves presents a characteristic Airy phase. For each array, time domain seismograms are stacked after correcting for dispersion and attenuation across the array, assuming plane wave propagation from an arbitrary azimuth (see Methods, and Supplementary Fig. 1). We first exercised and tested the array sensitivity on real earthquake data, for the two-year period 2000–01.

In order to attempt detection and location of non-earthquake sources of surface wave energy, it is necessary to first remove all intervals of time affected by earthquakes of moment magnitude $M_w > 5.5$ (see Methods), whereas smaller earthquakes do not significantly contribute to Rayleigh wave energy above 150 s (refs 1, 2). Many detections of Rayleigh wave energy are observed

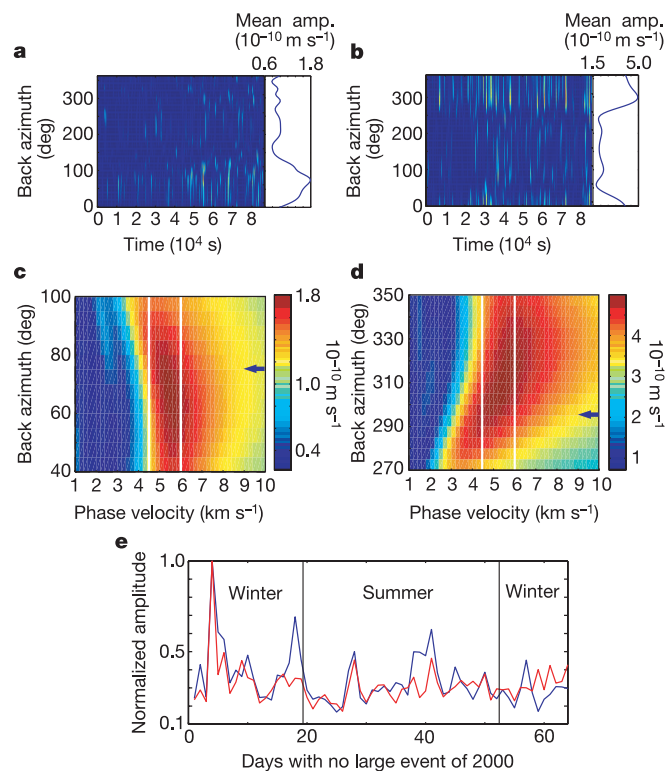


Figure 1 Analysis of detections for 31 January 2000. **a**, Amplitude of F-net stacks as a function of time and back azimuth. Small panel on right shows mean amplitude, as a function of back azimuth. A gaussian filter centred at 240 s has been applied to waveforms before stacking. **b**, Same as **a** for BDSN. **c**, Mean amplitude plot as a function of phase velocity and back azimuth for F-net, confirming that the observed energy corresponds to Rayleigh wave arrivals. The two vertical white lines indicate the range of phase velocities expected for Rayleigh waves between periods of 200 and 400 s. The theoretical phase velocity at 240 s is 4.85 km s^{-1} . Blue arrow indicates the back azimuth of the maximum mean amplitude shown in **a**. **d**, Same as **c** for BDSN. **e**, Mean amplitude of stacks as a function of time for quiet days for BDSN (red) and F-net (blue), for back azimuths of 295° and 65° respectively, in winter and back azimuths 105° and 235° in summer, normalized to the maximum amplitude for the entire time span. These azimuths correspond to the average direction of maximum amplitude in each season. The correlation coefficient between the two time series is 0.78.

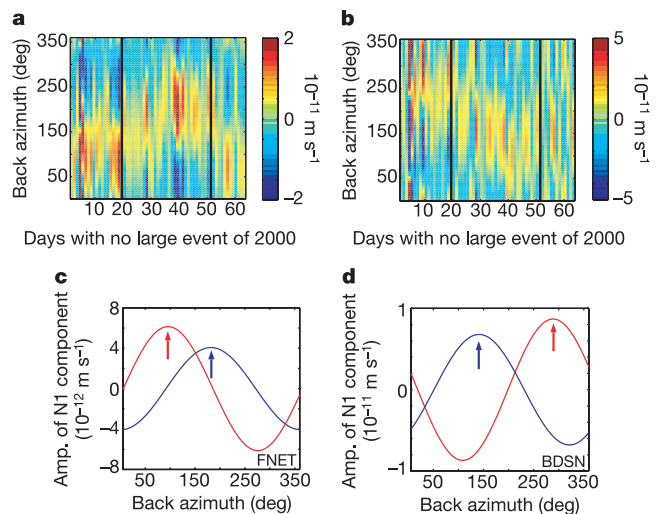


Figure 2 Amplitude of degree one as a function of time and back azimuth for the 64 quiet days in 2000. A 'quiet day' contains at least 12 contiguous hours uncontaminated by earthquakes, and only those intervals are considered within each day. **a**, Back azimuth corresponding to the maximum in the degree one component of stack amplitude for F-net as a function of time. Black vertical lines separate winter and summer intervals. Winter is defined as January–March, and October–December. **b**, Same as **a** for BDSN. **c**, Degree one as a function of azimuth for F-net, averaged for the winter (red) and the summer (blue). Arrows point to maxima in back azimuth; the degree one is almost as large as degree two in the data (see, for example, Supplementary Fig. 5) and is not contaminated by array response. **d**, Same as **c** for BDSN.

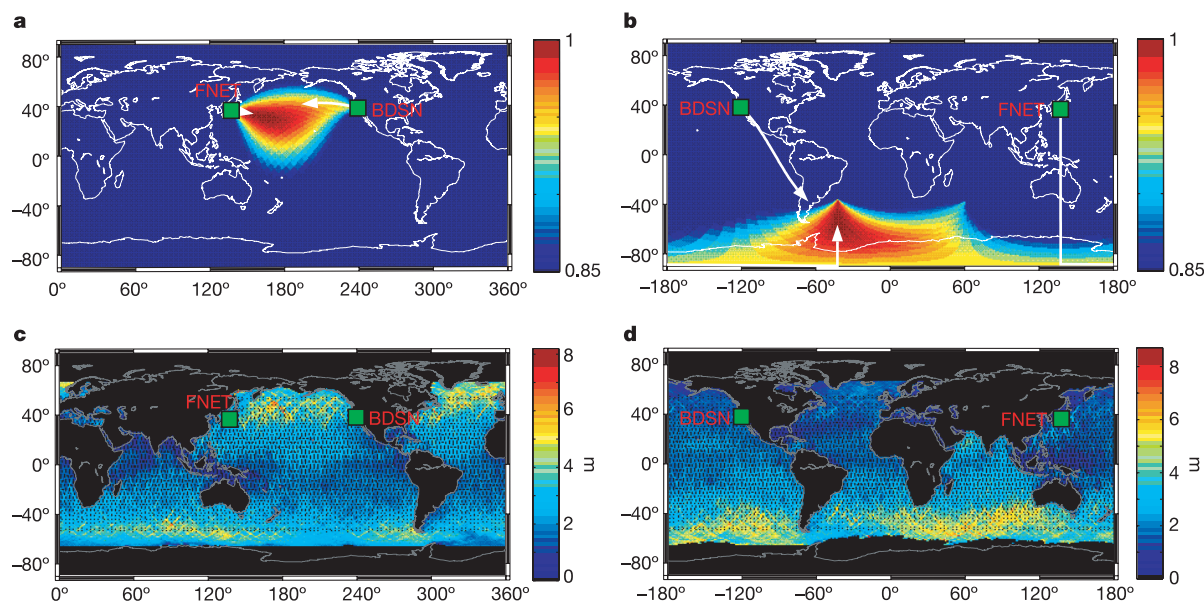


Figure 3 Comparison of seasonal variations in the distribution of hum-related noise (degree one only) and significant wave height in the year 2000. **a, b**, The directions corresponding to mean amplitudes that are larger than 85% of the maximum are combined for the two arrays in winter (**a**) and in summer (**b**) to obtain the region of predominant sources in each season. Arrows indicate the direction of maxima. Both

arrays are pointing to the North Pacific Ocean in the winter and to the southern oceans in the summer. **c, d**, Global distribution of significant wave height, in the winter (**c**) and in the summer (**d**), averaged from TOPEX/Poseidon images for the months of January and July 2000, respectively. Black colour in **c** and **d** indicates locations with no data.

on quiet days, unrelated to earthquakes (see, for example, Fig. 1a–d). Maximum stack amplitudes for BDSN and F-net correlate strongly as a function of time, indicating that the source of the Rayleigh wave ‘noise’ is common for the two arrays (Fig. 1e). Moreover, inspection of the Rayleigh wave energy distribution as a function of time and direction of arrival reveals striking spatial coherency. Because the array shape can introduce artificial distortions in the amplitude patterns as a function of azimuth, we first consider a component of the energy which is independent of the array response. For each day and for each array, we compute the Fourier spectrum of the stack amplitude as a function of azimuth and compare it to that of the array response (See Methods and Supplementary Information). The distribution of stack amplitudes as a function of time and back azimuth has a strong ‘degree one’ harmonic component in azimuth. This cannot be due to the array response which is dominated by even harmonics in azimuth, and indicates a true preferential direction of arrival. For each array, the direction of the maximum in degree one is stable at the seasonal scale; it is consistently different in (Northern Hemisphere) winter and summer (Fig. 2a, b). Then for each array, we compute the amplitude of the stack as a function of azimuth, averaged over summer and winter separately, calculate its Fourier spectrum and extract the degree one component (Fig. 2c, d) as well as the azimuth of the corresponding maximum. By combining the directions of maximum stack amplitude obtained for each array, in summer and winter respectively, and back-projecting along the corresponding great-circle paths, we infer that the sources locate preferentially in the northern Pacific Ocean in the winter (Fig. 3a), and in the southern oceans in the summer (Fig. 3b). These locations correspond to regions of maximum storm activity in the northern and southern winters respectively, as indicated by the comparison with significant wave height maps (Fig. 3c, d).

The degree one distribution only gives a first-order idea of the preferential direction of arrival at each array. More insight is obtained by determining how a distribution of sources, initially uniform in azimuth, needs to be modified to fit the original observed seasonal patterns. Such an analysis (See Methods, and Supplementary Figs 6 and 7) confirms that, in the winter, much of the energy originates in the North Pacific Ocean, while in the

summer, the activity shifts to the southern seas. We have also verified (through forward modelling experiments described in detail in Supplementary Information) that distributions of sources over continental areas are not compatible with the observations at both arrays simultaneously, whereas even a rough preferential distribution of sources in the northern Pacific fits the winter patterns for both arrays rather well. In the summer, we require a distribution of sources in the southern oceans, with preferential contributions from parts of the south Pacific and south Atlantic. Although a precise location of the sources would necessitate a more precise knowledge of the array response, our experiments clearly show that neither a uniform distribution of sources around the globe nor a distribution over continents are compatible with the data, in contrast to a distribution alternating between northern and southern oceans in winter and summer, respectively.

Finally, we considered a particular day, 31 January 2000, which corresponds to a maximum in the amplitude of the stacks for both F-Net and BDSN (Fig. 1e). We computed the stack amplitude as a function of azimuth over this 24 h interval at each array (Fig. 1a, b), and noted that the maximum average stack amplitude over this time interval points to a well defined direction of arrival, which is in general agreement with that found on other winter days. We confirm that the energy maxima correspond to the arrival of Rayleigh waves (Fig. 1c, d) by searching for the phase velocity and azimuth which give the maximum stack amplitude, averaged over that day. To analyse the time/space distribution of these sources, we conducted a parameter search in and around the north Pacific region. We find that the sources are distributed in space and time over a region spanning several thousand km², and with a correlation time of the order of ~6 h (Fig. 4).

Our results show that the ocean plays a key role in the excitation of the Earth’s ‘hum’. Part of the energy contained in ocean waves (generated by significant storms over the mid-latitude oceans) is converted to elastic waves. Infragravity waves are obvious candidates for the energy transfer from storms, through the ocean, to the sea floor¹⁹. They are indirectly driven by winds over ocean basins, and are probably generated in shallow water through conversion from short-period ocean waves by nonlinear processes²⁰. Some of

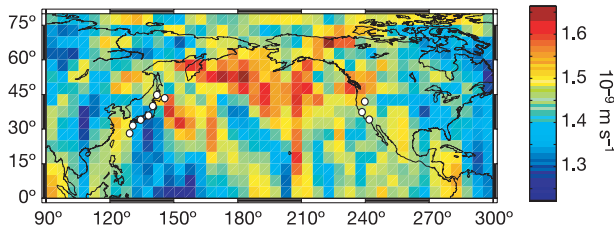


Figure 4 Distribution of sources for a 6 h time window on 31 January 2000 (from 14:00 to 20:00 UTC). Maximum amplitudes were averaged over all stations considered (in m s^{-1}), back-projected to the centre of $5^\circ \times 5^\circ$ blocks, and corrected for the dispersion of Rayleigh waves for each source station path, thereby indicating the distribution of possible source locations. F-net, BDSN (with 3 TERRAscope stations) and 10 European stations (KONO, ARU, BFO, DPC, KIEV, SSB, AQU, VSL, ESK and ISP) are used in this analysis. White dots in Japan and California indicate the areas spanned by the corresponding arrays. Waveforms have been band-pass filtered between 150 and 500 s. Analysis over shorter time windows does not lead to stable results, suggesting a correlation time of several hours and a spatially distributed source. In this analysis, the plane wave approximation is not made, nor are the results biased by the array response.

the energy leaks out and propagates as free waves into the ocean basins²¹. Hydrodynamic filtering may play a role in determining where the coupling occurs, in that it limits the frequency band in which infragravity waves interact with the deep ocean floor²², with shorter-period waves generated nearer to the coasts. The mechanism of generation of elastic waves probably involves focusing of infragravity waves by the concave shape of the continental boundaries towards the deep ocean, as well as by topography of the deep ocean floor. The efficiency with which long-period Rayleigh waves are generated in particular areas of the ocean basins must thus depend on the depth of the ocean floor, the shape of the continental shelves bounding the ocean basin, as well as the strength and persistence of storms. Notably, even though individual storms can be as strong over the north Atlantic as they are over the north Pacific, only the latter is detected in our data, which is in agreement with differences of 20–30 dB in pressure noise between these oceans in the infragravity wave band²³.

More detailed understanding of source distribution in space and time and its relation to ocean storms will require expanding the observational time windows by precisely removing signals due to earthquakes through a forward modelling approach, as is feasible at present with the availability of good quality three-dimensional elastic models of the mantle²⁴ as well as further analysis of wave height information. Also, the deployment of high-quality very broadband seismometers in the Southern Hemisphere (for example, across Australia) would help to characterize the source distribution better. Finally, we note that multi-disciplinary long-term ocean observatories spanning the entire water column (from sea floor to sea surface), such as are being proposed in the framework of the OOI (Ocean Observatories Initiative) programme²⁵, should acquire data that would help to make progress in our understanding of this complex energy transfer process. □

Methods

Array detection and location

We consider two regional networks of very broadband seismic stations in the Northern Hemisphere: (1) the Berkeley Digital Seismic Network (BDSN) in northern California, augmented by several stations of the TERRAscope network in southern California; and (2) the F-net in Japan. BDSN and F-net stations are equipped with very broadband STS-1 seismometers²⁶.

We pre-process vertical component time series by removing glitches and tides and deconvolving the instrument response to velocity. We then filter the data using either a gaussian filter centred at 240 s, or, in some experiments, a band-pass filter between 150 and 500 s (Fig. 4, Supplementary Fig. 4g). Two approximations are considered. We ignore the effect of lateral heterogeneity on the propagation of mantle Rayleigh waves, and we assume that they propagate across an array as plane waves (Supplementary Fig. 1), with a well defined incident azimuth (except in Fig. 4, where the plane wave approximation is not

used). The whole year is divided into 1-day intervals. For each time interval, and for each increment in back azimuth of 5° , we align the waveforms from 10 or more stations using the centre of the array as reference point, and correct for dispersion and attenuation of the fundamental mode Rayleigh wave across the array, according to the reference PREM velocity model²⁷. We stack the corrected waveforms using a phase weighted stack²⁸, which reduces uncorrelated noise, as compared to straight stacking (Supplementary Fig. 2). We tested the sensitivity of our array stacking for the detection of known earthquakes (Supplementary Information, Supplementary Figs 2, 3). Associating the detections obtained at three different arrays (including an array in western Europe) leads to an earthquake location method, not further described here, which works well at the magnitude 6 level, as we were able to detect and locate 85% of magnitude 6 and larger earthquakes without any optimization of the method.

Removal of intervals of time contaminated by earthquakes

We removed intervals of time of variable length, according to the size of the earthquake, for all earthquakes of magnitude $M_w > 5.5$. We first rejected any day within which an earthquake of $M_w > 6$ has occurred. We then defined a time span for additional rejection, using an exponential function constrained to be 3 days for $M_w \geq 7$, and 1.5 days for $M_w \geq 6$, starting at the origin time of the event. We also rejected a window before the earthquake, of length 10% of the length rejected following the origin time, to account for contamination due to the non-causal low-frequency filtering of the data. There are only a few 'quiet' windows left in one year, corresponding to $\sim 18\%$ of the time in 2000. Similar results are obtained for 2001.

Detection of Rayleigh waves during 'quiet' intervals

The same approach as used for detecting earthquakes is applied during quiet intervals. Supplementary Fig. 4 shows an example for a specific interval of time during day 2000/01/31, the same day as shown in Fig. 1. Alignment of traces is improved after taking into account Rayleigh wave dispersion in a specific azimuth range (Supplementary Fig. 4b, c). The corresponding stack amplitude has a maximum which is localized in time and azimuth (Supplementary Fig. 4a). The detected energy indeed corresponds to Rayleigh waves, as verified by a parameter search in azimuth and phase velocity space (Supplementary Fig. 4f) as well as in period versus phase velocity space (Supplementary Fig. 4g).

Removal of array response

Because the arrays considered are not completely symmetric, it is important to verify that biases due to the non-uniform array response in azimuth are not dominating the results. For this purpose, we have estimated the array response by randomly generating many synthetic seismograms and distributing them uniformly in azimuth. This procedure is described in detail in Supplementary Methods.

The amplitude of the array response cannot be directly compared to that of the observed patterns, so that we cannot completely remove it from the observed stack amplitudes. However, the array response has almost perfect 180° symmetry, with amplitude lobes corresponding to the elongated direction of the array, as expected. The corresponding Fourier spectra in azimuth are dominated by even degrees (particularly degree two), and have practically no degree one. In contrast, the spectra of the observed amplitude variations have a significant degree one (of a similar size as degree two, Supplementary Fig. 5e, f), which defines a specific direction of maximum amplitude, different in the summer from in the winter (Fig. 2). By conservatively analysing only the degree one component in the data, we guarantee that there is no contamination by the array response, while losing some directional resolution contained in the higher-degree azimuthal terms. Further experiments, as described in the main text and in Supplementary Information, confirm that the degree one analysis indeed gives a good indication of the spatial distribution of source energy and its seasonal variations.

Received 21 January; accepted 11 August 2004; doi:10.1038/nature02942.

- Suda, N., Nawa, K. & Fukao, Y. Earth's background free oscillations. *Science* **279**, 2089–2091 (1998).
- Tanimoto, T., Um, J., Nishida, K. & Kobayashi, N. Earth's continuous oscillations observed on seismically quiet days. *Geophys. Res. Lett.* **25**, 1553–1556 (1998).
- Tanimoto, T. & Um, J. Cause of continuous oscillations of the Earth. *J. Geophys. Res.* **104**, 28723–28739 (1999).
- Ekström, G. Time domain analysis of Earth's long-period background seismic radiation. *J. Geophys. Res.* **106**, 26483–26493 (2001).
- Beroza, G. & Jordan, T. H. Searching for slow and silent earthquakes using free oscillations. *J. Geophys. Res.* **95**, 2485–2510 (1990).
- Kobayashi, N. & Nishida, N. Continuous excitation of planetary free oscillations by atmospheric disturbances. *Nature* **395**, 357–360 (1998).
- Fukao, Y., Nishida, K., Suda, N., Nawa, K. & Kobayashi, N. A theory of the Earth's background free oscillations. *J. Geophys. Res.* **107**, doi:10.1029/2001JB000153 (2002).
- Wada, S. & Masters, G. Oceanic excitation of the continuous oscillations of the Earth. *Eos* **82**, F871 (2001).
- Nishida, K. & Kobayashi, N. Statistical features of Earth's continuous free oscillations. *J. Geophys. Res.* **104**, 28741–28750 (1999).
- Roult, G. & Crawford, W. Analysis of 'background' free oscillations and how to improve resolution by subtracting the atmospheric pressure signal. *Phys. Earth Planet. Inter.* **121**, 325–338 (2000).
- Rhie, J. & Romanowicz, B. Detection and location of potential sources of background low frequency surface wave energy. *Eos* **83** (Fall Meet. Suppl.), abstr. S12A–1184 (2002).
- von Seggern, D. Relative location of seismic events using surface waves. *Geophys. J. R. Astron. Soc.* **26**, 499–513 (1972).
- Rouland, D., Condis, C., Parmentier, C. & Souriau, A. Previously undetected earthquakes in the Southern Hemisphere from long-period Geoscope data. *Bull. Seismol. Soc. Am.* **82**, 2448–2463 (1992).

14. Rost, S. & Thomas, C. Array seismology: Method and applications. *Rev. Geophys.* **40**, doi:10.1029/2000RG000100 (2002).
15. Friedrich, A., Krüger, F. & Klinge, K. Ocean-generated microseismic noise located with the GRFO array. *J. Seismol.* **2**, 47–64 (1998).
16. Schulte-Pelkum, V., Earle, P. S. & Vernon, F. L. Strong directivity of ocean-generated seismic noise. *Geochem. Geophys. Geosyst.* **5**, doi:10.1029/2003GC000520 (2004).
17. Ekström, G., Nettles, M. & Abers, G. A. Glacial earthquakes. *Science* **302**, 622–624 (2003).
18. Nishida, K., Kobayashi, N. & Fukao, Y. Origin of Earth's ground noise from 2 to 20 mHz. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL013862 (2002).
19. Tanimoto, T. Jet stream, roaming ocean waves, and ringing Earth. *Eos* **84** (Fall Meet. Suppl.), abstr. S12F–04 (2003).
20. Webb, S., Zhang, X. & Crawford, W. Infragravity waves in the deep ocean. *J. Geophys. Res.* **96**, 2723–2736 (1991).
21. Munk, W., Snodgrass, F. & Gilbert, F. Long waves on the continental shelf: an experiment to separate trapped and leaking modes. *J. Fluid Mech.* **20**, 529–554 (1964).
22. Webb, S. & Crawford, W. Long-period seafloor seismology and deformation under ocean waves. *Bull. Seismol. Soc. Am.* **89**, 1535–1542 (1999).
23. Webb, S. Broadband seismology and noise under the ocean. *Rev. Geophys.* **36**, 105–142 (1995).
24. Ekström, G., Tromp, J. & Larson, E. Measurements and models of global surface wave propagation. *J. Geophys. Res.* **102**, 8137–8157 (1997).
25. Committee for Implementation of a Seafloor Observatory Network for Oceanographic Research. *Enabling Ocean Research in the 21st Century: Implementation of a Network of Ocean Observatories* (Ocean Studies Board, National Academies Press, Washington DC, 2003).
26. Wielandt, E. & Streckeisen, G. The leaf spring seismometer: design and performance. *Bull. Seismol. Soc. Am.* **72**, 2349–2367 (1982).
27. Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
28. Schimmel, M. & Paulsen, H. Noise detection and reduction of weak, coherent signals through phase-weighted stacks. *Geophys. J. Int.* **130**, 497–505 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the operators of the following seismic networks for making their data publicly available: BDSN (<http://www.seismo.berkeley.edu>), F-net (<http://www.fnet.bosai.go.jp/freesia/index.html>), IRIS (<http://www.iris.edu>) and TERRAscope. The Monthly Mean Global Surface Ocean Variables were obtained from the Physical Oceanography Distributed Active Archive Center (<http://podaac.jpl.nasa.gov/poet>). This work was partially supported by the NSF.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.R. (barbara@seismo.berkeley.edu).

Two new carnivores from an unusual late Tertiary forest biota in eastern North America

Steven C. Wallace¹ & Xiaoming Wang²

¹Department of Physics, Astronomy, and Geology, East Tennessee State University, Box 70636, Johnson City, Tennessee 37614, USA

²Department of Vertebrate Paleontology, Natural History Museum of Los Angeles County, 900 Exposition Boulevard, Los Angeles, California 90007, USA

Late Cenozoic terrestrial fossil records of North America are biased by a predominance of mid-latitude deposits, mostly in the western half of the continent. Consequently, the biological history of eastern North America, including the eastern deciduous forest, remains largely hidden. Unfortunately, vertebrate fossil sites from this vast region are rare^{1,2}, and few pertain to the critically important late Tertiary period, during which intensified global climatic changes took place^{3,4}. Moreover, strong phylogenetic affinities between the flora of eastern North America and eastern Asia clearly demonstrate formerly contiguous connections, but disparity among shared genera (eastern Asia–eastern North America disjunction) implies significant periods of separation since at least the Miocene epoch^{1,2}. Lacustrine sediments deposited within a former sinkhole in the

southern Appalachian Mountains provide a rare example of a late Miocene to early Pliocene terrestrial biota from a forested ecosystem⁵. Here we show that the vertebrate remains contained within this deposit represent a unique combination of North American and Eurasian taxa. A new genus and species of the red (lesser) panda (*Pristinailurus bristoli*), the earliest and most primitive so far known, was recovered. Also among the fauna are a new species of Eurasian badger (*Arctomeles dimolodontus*) and the largest concentration of fossil tapirs ever recorded. Cladistical analyses of the two new carnivores strongly suggest immigration events that were earlier than and distinct from previous records^{6,7}, and that the close faunal affinities between eastern North America and eastern Asia in the late Tertiary period are consistent with the contemporaneous botanical record^{8,9}.

The Gray Fossil Site consists of a sequence of finely laminated clays, silts and fine sands intermixed with isolated gravel lenses that fill a former sinkhole within the Cambrian/Ordovician Knox Group near the small community of Gray in Washington Co., Tennessee. The deposit covers roughly 1.8–2.0 ha, is up to 39 m thick and is the result of a small lake or pond that formed within the sinkhole. Subsequent weathering and erosion of the enclosing bedrock has generated a reversed topography, leaving the site as a high point on the landscape.

Vertebrate taxa such as *Tapiravus*, *Plionarctos*, *Pristinailurus* and *Arctomeles* (Table 1, left column) and abundant plant macrofossils (Table 1, right column) from *Quercus* (acorns) and *Carya* (hickory nuts) indicate that a dense forest surrounded the former 'pond'. *Quercus* and *Carya* constitute nearly 70% of initial pollen samples, and except for *Pinus* (which accounts for roughly 9% of the pollen count), the remaining taxa seem to be minor components of the flora. Both micro- and macrofossils reveal an arboreal flora, which was similar to that found in lower elevations of the southern Appalachians today.

The stratigraphic range of the rhino *Teleoceras*^{10,11} and the short-faced bear *Plionarctos*^{12,13} constrain the age of the assemblage to between 4.5 and 7 Myr (late Miocene to early Pliocene). This age is important because it occurs subsequent to the C₃/C₄ plant transition⁴; that is, when grasses first become dominant in many ecosystems worldwide¹⁴, leading to the prevalence of many grassland-adapted taxa (horses, camels, antilocaprids and so on) within other Miocene/Pliocene faunas (particularly in North America). However, the absence of these grassland-adapted taxa and the predominance of forest-adapted taxa (Table 1) suggest that the Gray Fossil Site may have acted as a refugium from the changing

Table 1 Vertebrate* and pollen taxa from the Gray Fossil Site

Vertebrates	Pollens
Reptilia	Conifers
<i>Trachemys</i> sp. ²⁰	<i>Pinus</i> (pine)
<i>Chrysemys</i> sp.	<i>Tsuga</i> (hemlock)
<i>Alligator</i> sp.	Deciduous
cf. <i>Sistrurus</i> sp. ²⁰	<i>Quercus</i> (oak)
cf. <i>Regina</i> sp. ²⁰	<i>Carya</i> (hickory)
Aves	<i>Ulmus</i> (elm)
Passeriformes	<i>Betula</i> (birch)
Mammalia	<i>Fraxinus</i> (ash)
Soricidae	<i>Celtis</i> (hackberry)
Rodentia	Shrubs
Gomphotheriidae	<i>Alnus</i> (alder)
<i>Tapiravus polkensis</i>	<i>Salix</i> (willow)
<i>Teleoceras</i> sp.	Herbs
Tayassuidae	<i>Ambrosia</i> -type ('ragweed')
cf. <i>Megatylopus</i> sp.	Cyperaceae (sedge)
cf. <i>Machairodus</i> sp.	Gramineae (grass)
<i>Plionarctos</i> sp.	Umbelliferae (parsely family)
Canidae	Caryophyllaceae (pink family)
<i>Pristinailurus bristoli</i> gen. et sp. nov.	
<i>Arctomeles dimolodontus</i> sp. nov.	
Excluding fishes and amphibians.	

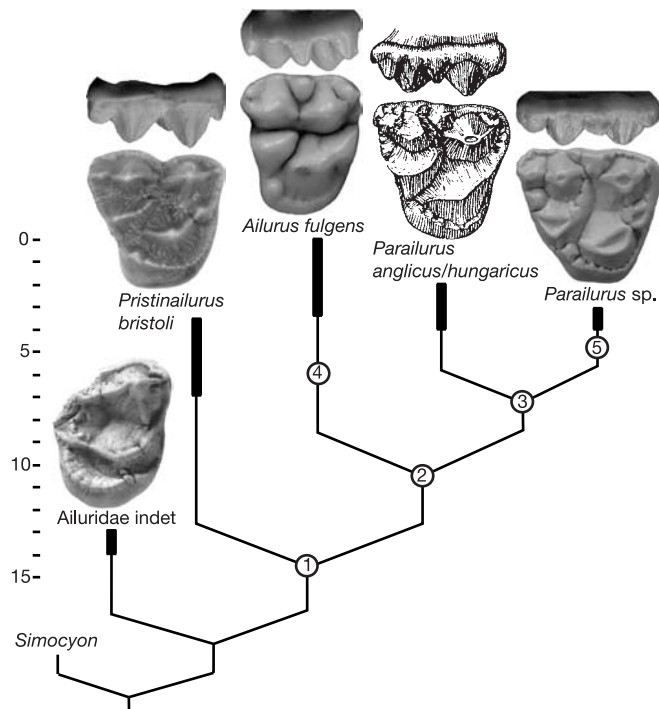


Figure 1 Phylogenetic relationship and geological ages of red panda fossils. Taxa illustrated are: Ailuridae indet., FSL 66113, Four locality, France, middle Miocene²¹; *Pristinailurus bristoli* gen. et sp. nov., ETMNH-360, Gray Fossil Site, Tennessee, late Miocene/early Pliocene; *Ailurus fulgens*, LACM 62839, San Diego zoo, late Pliocene²² to present; *Parailurus anglicus/hungaricus*, SMF 2000/235 (= Wö 21 in Tedford⁹), Wölferstheim, Germany, early Pliocene to late Pliocene^{6,23–26}, and *Parailurus* sp., LACM 10808, White Bluffs, Washington, early Pliocene⁶. We use *Simocyon* as its nearest outgroup for the family²⁷. M¹s are scaled to be approximately equal in size. Scale represents age (Myr before present).

environment of the late Miocene, because of its isolation from the more 'open' ecosystems typical of similar-aged fossil deposits.

A unique combination of taxa makes the Gray Fossil Site very distinct from other late Miocene and early Pliocene faunas. Not only is there a strong southern influence (warm adapted) indicated by the occurrence of *Alligator* and *Tapiravus*, but also there is a strong northern component (cool adapted) represented by taxa such as *Arctomeles* and *Pristinailurus*. Just as important are the conspicuous absence of equids, which are typically abundant in Miocene and Pliocene faunas of central and western North America (often represented by multiple taxa)¹⁵, and the overabundance of tapirs, which are typically only minor components of faunas¹⁶. Although younger than the peak of equid diversity, between 15–8 Myr ago¹⁵, the lack of equids (or other such cursorial taxa) suggests little influence from the Great Plains or Gulf Coast grassland environments. Furthermore, the prolific number of individual tapirs at the site supports the dominance of an arboreal setting (at least in the immediate proximity of the sinkhole). Though *Teleoceras* has been considered a grazer¹⁷, recent isotopic analysis¹⁸ has shown that at least some taxa were mixed feeders, which would not preclude its presence in a forested ecosystem.

Terrestrial carnivores are frequently long-distance dispersers because of their predatory behaviours. Consequently, their sudden appearance on different continents often serves as a useful marker for mammalian biochronology¹⁹. Therefore, in addition to revealing possible dispersal corridors previously hidden from our view because of the biased North American terrestrial fossil record, these new 'exotic' hypocarnivores (carnivores that specialize in highly omnivorous diets) from the Gray Fossil Site will be ideal biostratigraphic markers where recovered.

Living red pandas are confined to the mountainous belt of the southern Himalayas, and like the giant panda, have a highly specialized diet of bamboo. This herbivorous diet is a rare occurrence among carnivores. Although there is no evidence for the presence of bamboo in the Gray site (possibly owing to inherent difficulties in pollen identification in the Poaceae), it seems

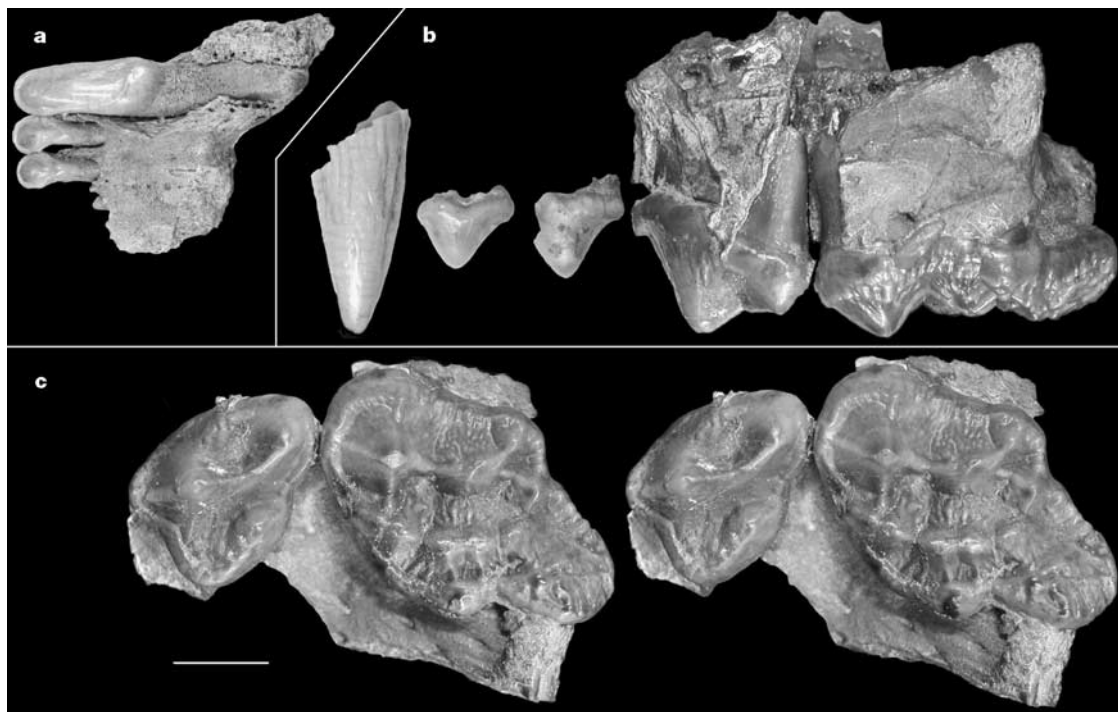


Figure 2 *Arctomeles dimolodontus* sp. nov., ETMNH-361, holotype. **a**, Occlusal view of left premaxillary fragment with I^{1–3}. **b**, Lateral view of right C (reversed) and left P²–M¹. **c**, Occlusal view of left maxillary with P⁴ and M¹ (stereo pair). Scale bar, 5 mm.

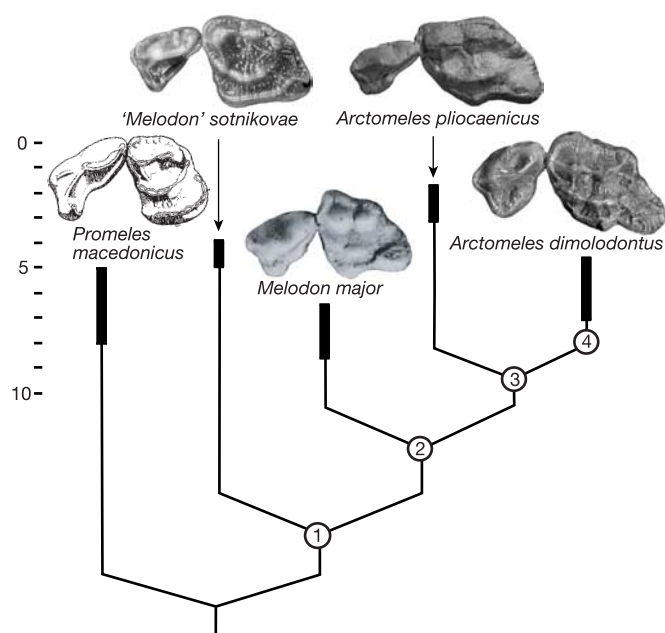


Figure 3 Phylogenetic relationship and geological ages of the Gray-site badger. Taxa illustrated are: *Promeles macedonicus* (reversed from the right side), MA 406, Maramena, Greece, late Miocene²⁸; “*Melodon*” *sotnikovae*, CMN 51770, Ellesmere Island, Canada, early Pliocene⁷; *Melodon major*, locality 49, Yangmugou, Baode, Shanxi Province, China, late Miocene²⁹; *Arctomeles pliocaenicus*, MF338, Weze 1 locality, Poland, early Pliocene³⁰; and *Arctomeles dimolodontus* sp. nov., ETMNH-361, Gray Fossil Site, Tennessee, late Miocene/early Pliocene. We also include *Promeles*, a primitive meline, as an outgroup to establish character polarity^{28,29}. Scale represents age (Myr before present).

probable that river cane (*Arundinaria gigantea*), a member of the bamboo clade that is native to Tennessee, or a similar form, may have been more widespread in the late Miocene. However, it is likely that the Gray-site panda was able to subsist on non-bamboo leaves while passing through an arctic arboreal corridor and ultimately found more habitable land in the eastern deciduous forest of the southern Appalachians.

Both new hypocarnivores (*Pristinailurus bristoli* and *Arctomeles dimolodontus*) show unique phylogenetic relationships with Eurasian taxa, indicating both the presence of a strong connection (physical and genetic) between the two regions until at least the late Miocene, and the likelihood of multiple immigration events into the New World from different Eurasian stock. The relationships and timing of these events are consistent with floral eastern Asia–eastern North America disjunction^{1,2}.

Methods

Pristinailurus bristoli gen. et sp. nov.

Etymology. *Pristinus* (Latin), meaning former or previous, and *Ailurus* for the living genus. Species after Larry Bristol who found the holotype. Common name: Bristol’s Appalachian panda.

Holotype. Right first upper molar (RM¹) (East Tennessee Museum of Natural History = ETMNH-360); measurements (labial length × transverse width): 11.4 × 12.3.

Referred material. Right first upper canine (RC¹) (ETMNH-359) and RM¹ (ETMNH-360) (Fig. 1).

Diagnosis. *Pristinailurus* differs from *Parailurus* and *Ailurus* in its retention of primitive characters on M¹: length less than width; metacone enlarged, strongly crest-like and isolated from the protocone; metastylar and parastylar cusps poorly developed to lacking; absence of both a mesostylar cusp and a discrete accessory cusp on posterior cingulum between metastyle and metacone; retention of a lingual cingulum that merges with the hypocone at its posterior end.

Discussion. Tedford and Gustafson⁶ reported *Parailurus* from the early Blancan of Washington on the basis of a single RM¹. Fortunately the same tooth was collected from the Gray site, allowing direct comparison (Fig. 1). Shared derived characters at numbered nodes in Fig. 1 are as follows: (1) enlarged M¹ metacone that is strongly crest-like and isolated from the protocone, presence of a metastylar cusp; (2) presence of a distinct mesostylar cusp, reduced lingual cingulum, a discrete hypocone on posterior end of

lingual cingulum; (3) M¹ length equal to or greater than width, presence of a discrete accessory cusp on posterior cingulum between metastyle and metacone; (4) presence of a distinct and enlarged parastylar cusp, major cusps swollen at the bases to encroach valleys between cusps; and (5) M¹ length greater than its width, enlargement of a discrete protocone (see further explanation in Supplementary Information).

The substantially more primitive Gray panda, in contrast to a highly derived *Parailurus* from Washington, permits evaluation of character polarities that were previously obscure. Our new phylogeny (Fig. 1) shows that *Pristinailurus* is at least five steps (characters) removed from the Washington *Parailurus*, and that the two North American pandas are separated phylogenetically by two intermediate Eurasian forms. Such a topology strongly suggests a separate immigration event independent from that represented by the Washington *Parailurus*, a pattern also consistent with the age estimate for the Gray site, which is older than the Taunton site in Washington.

Arctomeles Stach, 1951

Arctomeles dimolodontus sp. nov.

Etymology. Based on the inferred occlusion of the talon-like basin on M¹ with the second lower molar (M₂) (see Fig. 2). *di* (Greek) meaning two; *mola* (Latin) for millstone (as in molar); *odontos* (Greek) denoting tooth. Literally meaning ‘two millstone tooth’ or ‘two molar tooth’. Common name: woodland badger.

Holotype. ETMNH-361 left premaxillary fragment with incisors LI^{1–3}; associated premolars LP^{2–4}, molar LM¹, incisors RI^{1–3} and canine RC¹ (Fig. 2); measurements (labial length × transverse width): P⁴, 11.8 × 8.6 mm; M¹, 12.2 (maximum anteroposterior length: 18.3 mm) × 12.2 mm.

Diagnosis. Similar in size to *Arctomeles pliocaenicus*. P⁴ has an expanded cingular shelf and is large relative to M¹; P⁴ hypocone is larger, and more posterior than in *A. pliocaenicus*; M¹ lingual cingulum expands more anterolaterally than in *A. pliocaenicus* but not nearly as far as in “*Melodon*” *sotnikovae*; talon-like basin on M¹ extends well posterior of metacone and hypocone; however, there is a distinct constriction between the basin and the remainder of the tooth; postprotocrista of M¹ is divided into three discrete cusps; M¹ has a metastylar cusp and enlarged paracone; accessory cusps and crenulations within the basins on both the P⁴ and M¹ are larger and more elaborate than in *A. pliocaenicus*.

Included species. Listed in Tedford and Harington⁷ except “*Melodon*” *sotnikovae*.

Discussion. This new taxon is more derived than other members of the genus in the development of the talon-like basin on M¹, size and morphology of the P⁴, and complexity of accessory cusps and crenulations on all teeth (Fig. 3). Shared derived characters at numbered nodes in Fig. 3 are as follows: (1) expansion of posterior part of talon on M¹, M¹ postprotocrista oriented anteroposteriorly resulting in a posteriorly open talon to occlude with a long talonid basin of M₁, presence of multiple cusps along posterior cingulum and trailing behind metacone, and presence of metastylid on M₁; (2) M¹ posterior talon curves dorsally to be in contact with an expanded M₂, and presence of a P⁴ hypocone; (3) P⁴ protocone reduced to a low crest, and reduction of anterior portion of M¹ lingual cingulum; and (4) postprotocrista divided into three discrete cusps, presence of a M¹ metastylar cusp, and enlargement of P⁴ hypocone (see further explanation in Supplementary Information).

Received 15 April; accepted 8 July 2004; doi:10.1038/nature02819.

- Emry, R. J. & Eshelman, R. E. in *Geology and Paleontology of the Lower Miocene Pollack Farm Fossil Site, Delaware* (ed. Benson, R. N.) 153–173 (Delaware Geological Survey special publication, Newark, 1998).
- Farlow, J. O. et al. The Pipe Creek Sinkhole biota, a diverse late Tertiary continental fossil assemblage from Grant County, Indiana. *Am. Midl. Nat.* **145**, 367–378 (2001).
- Pagani, M., Freeman, K. H. & Arthur, M. A. Late Miocene atmospheric CO₂ concentrations and the expansion of C₄ grasses. *Science* **285**, 876–879 (1999).
- Cerling, T. E. et al. Global vegetation change through the Miocene/Pliocene boundary. *Nature* **389**, 153–158 (1997).
- Wallace, S. C., Nave, J. W. & Burdick, K. M. Preliminary report on the recently discovered Gray Fossil Site (Miocene), Washington Co., Tennessee: with comments on observed paleopathologies—The advantages of a large sample. *J. Vert. Paleontol.* **22**, 117 (2002).
- Tedford, R. H. & Gustafson, E. P. First North American record of the extinct panda *Parailurus*. *Nature* **265**, 621–623 (1977).
- Tedford, R. H. & Harington, C. R. An Arctic mammal fauna from the Early Pliocene of North America. *Nature* **425**, 388–390 (2003).
- Wen, J. Evolution of eastern Asian and eastern North American disjunct distributions in flowering plants. *Annu. Rev. Ecol. Syst.* **30**, 421–455 (1999).
- Qian, H. & Ricklefs, R. E. Large-scale processes and the Asian bias in species diversity of temperate plants. *Nature* **407**, 180–182 (2000).
- Prothero, D. R., Guerin, D. R. & Manning, E. in *The Evolution of the Perissodactyls* (eds Prothero, D. R. & Schoch, R. M.) 320–340 (Oxford Univ. Press, New York, 1989).
- Prothero, D. R. in *Evolution of Tertiary Mammals of North America* Vol. 1 (eds Janis, C. M., Scott, K. M. & Jacobs, L. L.) 595–605 (Cambridge Univ. Press, Cambridge, 1998).
- Hunt, R. M. Jr in *Evolution of Tertiary Mammals of North America* Vol. 1 (eds Janis, C. M., Scott, K. M. & Jacobs, L. L.) 174–195 (Cambridge Univ. Press, Cambridge, 1998).
- Tedford, R. H. & Martin, J. *Plionarctos*, a tremarctine bear (Ursidae; Carnivora) from western North America. *J. Vert. Paleontol.* **21**, 311–321 (2001).
- Jacobs, B. F., Kingston, J. D. & Jacobs, L. L. The origin of grass-dominated ecosystems. *Ann. Mo. Bot. Gard.* **86**, 590–643 (1999).
- MacFadden, B. J. & Hulbert, R. C. J. Explosive speciation at the base of the adaptive radiation of Miocene grazing horses. *Nature* **336**, 466–468 (1988).
- Hulbert, R. C. J. A new *Tapirus* (Mammalia, Perissodactyla) from the late Miocene of Florida. *Bull. Fla. Mus. Nat. Hist.* (in the press).
- Voorhies, M. R. & Thomasson, J. R. Fossil grass anthecia within Miocene rhinoceros skeletons: diet in an extinct species. *Science* **206**, 331–333 (1979).
- MacFadden, B. J. Tale of two rhinos: isotopic ecology, paleodiet, and niche differentiation of *Aphelops* and *Teleoceras* from the Florida Neogene. *Paleobiology* **24**, 274–286 (1998).

19. Tedford, R. H., et al. in *Late Cretaceous and Cenozoic Mammals of North America* (ed. Woodburne, M. O.) 169–231 (Columbia Univ. Press, New York, 2004).
20. Ginsburg, L., Maridet, O. & Mein, P. Un Ailurinae (Mammalia, Carnivora, Ailuridae) dans le Miocène moyen de Four (Isère, France). *Geodiversitas* **23**, 81–85 (2001).
21. Kundrát, M. New dental remains of an extinct lesser panda—morphotype or new species? *J. Vert. Paleontol.* **17**, 58A (1997).
22. Boyd-Dawkins, W. On *Ailurus anglicus*, a new carnivore from Red Crag. *Q. J. Geol. Soc. Lond.* **44**, 228–231 (1888).
23. Kormos, T. H. Beiträge zur Kenntnis der Gattung *Parailurus*. *Mitt. Jahrb. K. Ung. Geol. Anst.* **30**, 1–39 (1935).
24. Schlosser, M. *Parailurus anglicus* und *Ursus böckhi*, aus den Ligniten von Baróth-Köpecz, Comitat Hármezsek in Ungarn. *Mitt. Jahrb. K. Ung. Geol. Anst.* **13**, 66–95 (1899).
25. Morlo, M. & Kundrát, M. The first carnivore fauna from the Ruscium (Early Pliocene, MN 15) of Germany. *Paläontol. Z.* **75**, 163–187 (2001).
26. Wang, X. New cranial material of *Simocyon* from China, and its implications for phylogenetic relationship to the red panda (*Ailurus*). *J. Vert. Paleontol.* **17**, 184–198 (1997).
27. Schmidt-Kittler, N. in *The Vertebrate Locality Maramena (Macedonia, Greece) at the Turolian-Ruscinian Boundary (Neogene)* (ed. Schmidt-Kittler, N.) 75–86 (Münchner Geowissenschaftliche Abhandlungen, Munich, 1995).
28. Zdansky, O. Jungtertiäre Carnivoren Chinas. *Palaeont. Sin.* **C 2**, 1–149 (1924).
29. Stach, J. *Arctomeles pliocenicus*, nowy rodzaj i gatunek z podrodziny borsukowatych (*Arctomeles pliocenicus* nov. gen. & sp. from Weze). *Acta Geol. Pol.* **2**, 129–157 (1951).
30. Parmalee, P. W., Klippel, W. E., Meylan, P. A. & Holman, J. A. A late Miocene–early Pliocene population of *Trachemys* (Testudines: Emydidae) from east Tennessee. *Ann. Carnegie Mus.* **71**, 233–239 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements S.C.W. wishes to thank former Tennessee Governor Don Sundquist for protecting the fossil deposit and awarding significant funds for its preservation and study; the Tennessee Department of Transportation, East Tennessee State University, Office of Research and Sponsored Programs, and the College of Arts and Sciences for their continuing support of this project; and Larry Bristol for discovering the *Pristinailurus* M¹ and bringing it to my attention. X.W. wishes to acknowledge the Chinese National Natural Science Foundation and the National Geographic Society for support in comparative studies.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.W. (xwang@nhm.org).

New evidence on the earliest human presence at high northern latitudes in northeast Asia

R. X. Zhu¹, R. Potts², F. Xie³, K. A. Hoffman⁴, C. L. Deng¹, C. D. Shi¹, Y. X. Pan¹, H. Q. Wang¹, R. P. Shi¹, Y. C. Wang¹, G. H. Shi¹ & N. Q. Wu¹

¹Paleomagnetism Laboratory, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China

²Human Origins Program, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0112, USA

³Hebei Province Institute of Cultural Relics, Shijiazhuang 050000, China

⁴Physics Department, California Polytechnic State University, San Luis Obispo, California 93410, USA

The timing of early human dispersal to Asia is a central issue in the study of human evolution. Excavations in predominantly lacustrine sediments at Majuangou, Nihewan basin, north China, uncovered four layers of indisputable hominin stone tools. Here we report magnetostratigraphic results that constrain the age of the four artefact layers to an interval of nearly 340,000 yr between the Olduvai subchron and the Cobb Mountain event. The lowest layer, about 1.66 million years old (Myr), provides the oldest record of stone-tool processing of animal tissues in east Asia. The highest layer, at about 1.32 Myr, correlates with the stone tool layer at Xiaochangliang¹, previously considered the oldest archaeological site in this region. The findings at Majuangou indicate that the oldest known human presence in northeast Asia

at 40° N is only slightly younger than that in western Asia^{2,3}. This result implies that a long yet rapid migration from Africa, possibly initiated during a phase of warm climate, enabled early human populations to inhabit northern latitudes of east Asia over a prolonged period.

The Majuangou (MJG; 40° 13.517' N, 114° 39.844' E) section lies in the eastern margin of the Nihewan basin (Fig. 1). It is a lacustrine sequence with brief intervals of wetland and lake-margin sediments, and consists mainly of greyish-yellow and greyish-green clay, silty clay and silt. It is underlain by red Jurassic volcanic breccia. Loess sediments at the top of the section have been subjected to erosion. The four artefact layers found in the MJG section are, from top to bottom, Banshan⁴ (44.3–45.0 m), MJG-I (ref. 5; 65.0–65.5 m), MJG-II (73.2–73.56 m) and MJG-III (75.0–75.5 m) (Fig. 2).

The Banshan artefact layer, discovered and excavated in 1990 (2 m² area, 70 cm thick), contained 95 stone artefacts in gravelly sandy silt⁴. Excavation of MJG-I in 1993 (20 m², 50 cm) yielded 111 stone tools in clayey silt⁵. Renewed excavation at Majuangou in 2001 and 2002 uncovered 226 artefacts in brown clayey silt of MJG-II (40 m², 36 cm) and 443 artefacts in greyish-black silty clay of MJG-III (85 m², 50 cm). The sediments, numerous molluscan shells (*Gyraulus chihliensis* and *Planorbis youngi*), and leaves and fruits of aquatic plants (for example *Trapa* sp.) in MJG-III indicate a low-energy lakeshore or marsh environment rich in organic materials. The *in situ* artefact density in this layer was low overall (10.4 artefacts per m³), but artefacts and fauna in some 5-cm-thick

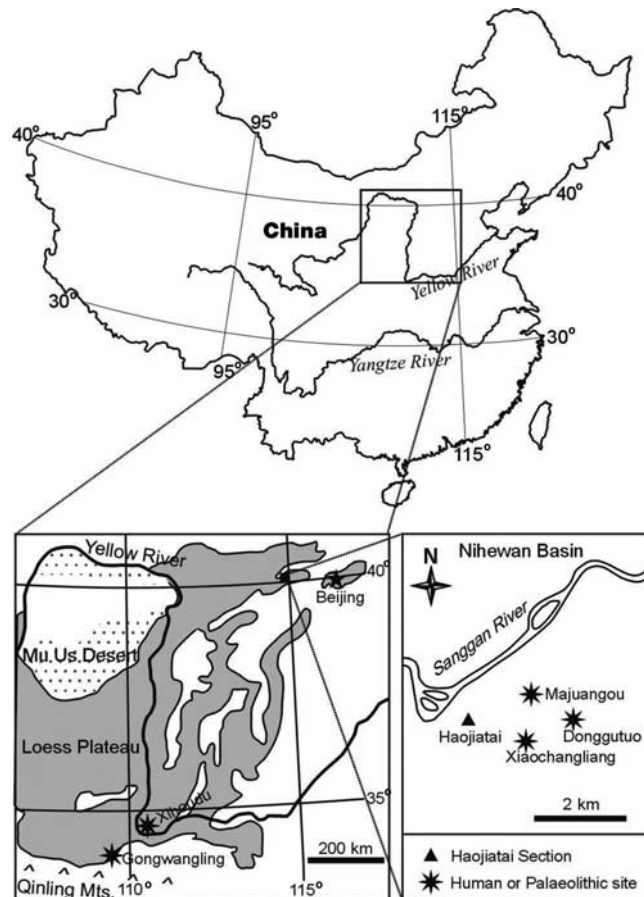


Figure 1 Location of the Majuangou and Haojiatai sections in the Nihewan basin. Some sites mentioned in the text, Xiaochangliang, Donggutuo, Gongwangling and Xihouzi, are indicated. The Qinling Mountains (bottom left) are the traditional dividing line between north and south China. The Yellow River and Yangtze River are the major river systems in north and south China, respectively.

units were as high as 170 specimens per m³ over the entire excavation and 620 specimens per m³ in a single square metre. These concentrations are comparable to those in African Plio-Pleistocene archaeological sites^{6,7}. MJG-III exhibits remarkable preservation demonstrated by animal-trampled sedimentary surfaces, very fresh condition of the artefacts, and fossil bone surface details that include tool percussion marks and numerous fine scratches attributed to trampling.

The four Majuangou layers preserve indisputable stone tools indicative of repetitive stone-on-stone percussion flaking (Fig. 3a–e). The assemblages are dominated by core fragments that exhibit truncated negative scars and by flakes with percussion platforms and bulbs. Each artefact layer also contained flaked cores that show

striking platforms and multiple overlapping negative scars. The cores can be placed in artefact categories of chopper, scraper and polyhedron also known in African Plio-Pleistocene stone tool assemblages. The MJG cores were chipped from angular fragments of chert, sandstone, quartz and andesite, and thus differ from typical East African Oldowan artefacts made on rounded lava cobbles. The artefacts of MJG-I to MJG-III are significantly outsized clasts in very fine-grain depositional contexts, which indicates the hominin transport of rocks from outcrop sources over an unknown distance.

Vertebrate fossil remains were best represented at MJG-III ($N = 1,014$), most of which are attributable to *Elephas* sp. Other taxa include horse *Equus sanmeniensis*, hyena *Pachycrocuta* sp., rhinoceros *Coelodonta antiquitatis*, deer *Cervus* sp., bovid *Gazella*

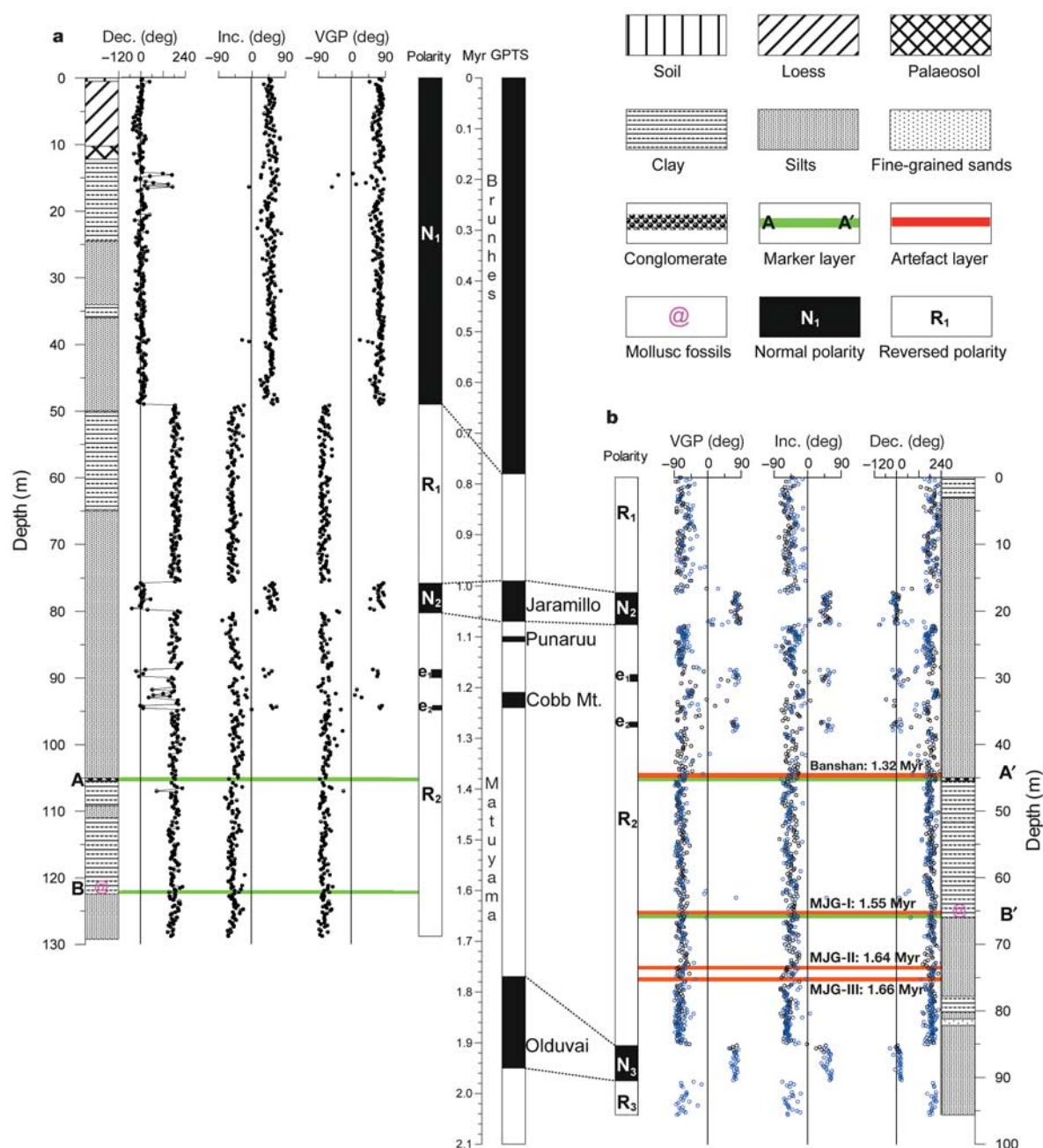


Figure 2 Lithostratigraphy and magnetostratigraphy and correlation with the geomagnetic polarity timescale (GPTS)⁸. **a**, Haojietai; **b**, Majuangou. The four artefact levels are shown. To confirm the palaeomagnetic results, two sets of parallel samples

(black and blue circles in **b**) with independent orientation were measured on the Majuangou outcrop and well samples. Inc., inclination; Dec., declination; VGP, virtual geomagnetic pole.

sp., ostrich *Struthio* sp., and *Carnivora* gen. et sp. indet. The mammals from MJG-III and the Banshan layer are typical of the taxa recorded in the Xiaochangliang site⁸. Evidence of sedimentary abrasion due to trampling hinders an unambiguous identification of purposeful tool butchery marks at MJG; however, several diaphysis fragments of deer- and horse-sized mammalian long bones show tool percussion damage indicative of marrow extraction (Fig. 3f). Although there was accumulation of tools and fossil bones during depositional hiatuses, there is no evidence of deflation surfaces that might have associated objects from separate strata. Accumulation of both artefacts and fossil animals was therefore contemporaneous, and the presence of tool-modified bones implies that hominins acquired food from the animal remains preserved in the MJG layers.

We examined the 95.6-m-thick MJG section palaeomagnetically and compared it with a 128.8-m-thick parallel section 1.5 km away named Haojiatai (HJT; 40° 13.240' N, 114° 38.938' E) (Fig. 1). The HJT section, which consists of flat-lying beds exposed in deep gullies, preserves the entire upper part of the Nihewan sequence, including the Holocene soil and the last glacial loess (Fig. 2). All HJT samples were collected from natural outcrops. Samples from the top of the MJG section to a depth of 75.2 m came from natural outcrops; two wells were dug to extend the MJG palaeomagnetic record. The first well, about 20 m southeast of the MJG-III site, recorded a depth interval from 75.2 to 86.2 m; the second well, about 50 m northwest of the site, recorded a depth interval from 75.2 to 95.6 m. The

sedimentary sequences at MJG and HJT are well correlated by two distinctive marker layers: a conglomerate layer (found at the 45-m depth at MJG and the 105-m depth at HJT) and a greyish-yellow clay layer with molluscan fossils (found at the 66-m depth at MJG and the 122.4-m depth at HJT) (Fig. 2).

Rock magnetic methods, which included anisotropy of magnetic susceptibility, thermomagnetic analysis and hysteresis measurements, showed that magnetite of pseudo-single-domain grain size is the principal carrier of the magnetic remanence and that the sedimentary magnetic fabric had been unperturbed since deposition. After this check on the reliability of the two palaeomagnetic records, we established the polarity stratigraphy through stepwise demagnetization of the natural remanent magnetization. Complete information on rock magnetic methods and demagnetization of the natural remanent magnetization used here is given in Supplementary Information.

After the removal of secondary remanent magnetization components from each sample through thermal and/or alternating field demagnetization procedures, virtual geomagnetic pole latitudes were determined from the characteristic remanent magnetization vector directions. These virtual geomagnetic poles were subsequently used to define the succession of magnetostratigraphic polarity in the two sections (Fig. 2).

Four magnetostratigraphic zones are recognized in the HJT section: two with normal polarity, N1 (0–49.0 m) and N2 (75.8–80.2 m); and two with reverse polarity, R1 (49.0–75.8 m) and R2 (80.2–128.8 m). In the MJG section there are five magnetostratigraphic zones: two normal and three reverse. These magnetostratigraphic zones correlate to the polarity sequence at HJT as follows: N2 (17.2–22.0 m) and N3 (85.0–90.5 m); R1 (0–17.2 m), R2 (22.0–85.0 m) and R3 (90.5–95.6 m). The sediment layers containing stone artefacts all occur within magnetozone R2 at MJG.

Because the Holocene soil, the last glacial loess, and soil associated with the last interglacial overlay the HJT lacustrine sequence, the magnetostratigraphic zones determined for HJT can readily be correlated to the geomagnetic polarity timescale⁹. HJT magnetostratigraphic zones N1 and N2 correspond to the Brunhes chron and the Jaramillo subchron, respectively; thus, magnetostratigraphic zones N2 and N3 in the MJG section correspond to the Jaramillo subchron and the Olduvai subchron, respectively. Hence, these Nihewan basin sediments were deposited from just before the onset of the Olduvai subchron into the Brunhes normal chron.

The magnetostratigraphic correlation is strengthened in that the mammalian fauna from the Banshan and MJG-III layers is late Pliocene to early Pleistocene in age^{8,10–12}. In addition, two short intervals of possible transitional field behaviour, labelled e1 and e2 in Fig. 2, are recorded within magnetozone R2 at both MJG and HJT (e1, 29.5–30.5 m at MJG and 88.7–89.9 m at HJT; e2, 36.5–37.3 m at MJG and 94.1–94.7 m at HJT). Given that the duration of magnetozone R2 is about 0.70 Myr—between the termination of the Olduvai subchron (1.77 Myr) and the onset of the Jaramillo subchron (1.07 Myr)⁹—the interpolated ages for e1 and e2 are 1.16 Myr and 1.24 Myr, respectively, based on an averaged rate of sediment deposition. These values are remarkably similar to the ⁴⁰Ar–³⁹Ar age determinations of 1.10–1.11 Myr (ref. 13) and 1.21–1.24 Myr (ref. 9) for the Punaruu and Cobb Mountain geomagnetic events. It is therefore possible that the sediments in both sections record not only the coarse magnetostratigraphy of the Matuyama chron (that is, the Jaramillo and Olduvai normal polarity subchrons) but also some of its fine structure.

The Banshan, MJG-I, MJG-II and MJG-III artefact layers within magnetozone R2, reflecting brief episodes of wetlands or lake-margin deposition within a largely lacustrine sequence, have mid-way depths of 44.65, 65.25, 73.38 and 75.25 m, respectively (Fig. 2). Again with the use of an averaged sediment accumulation rate for magnetozone R2 at MJG, the ages of these four artefact layers can be estimated at 1.32, 1.55, 1.64 and 1.66 Myr, respectively.

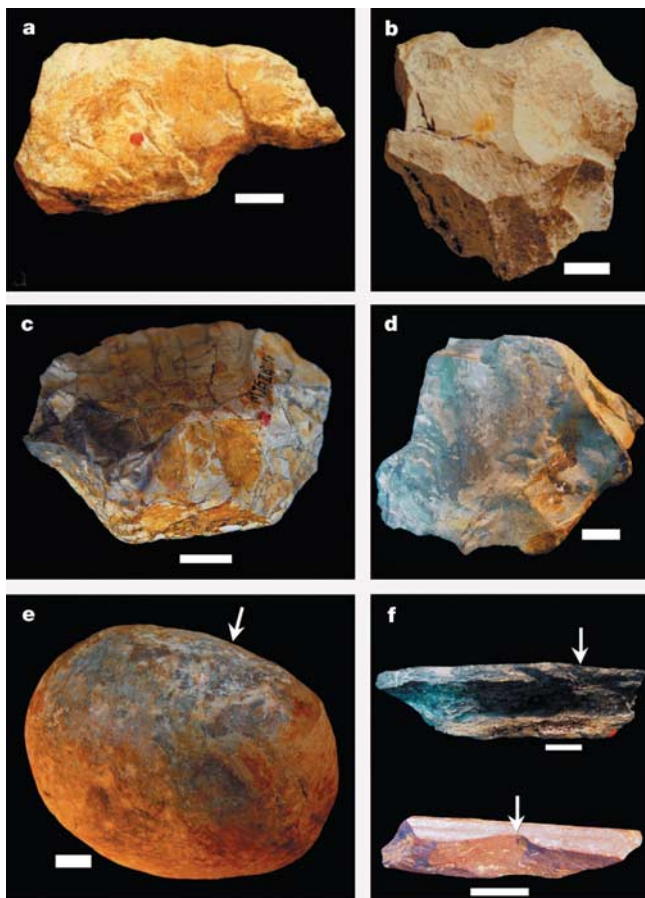


Figure 3 Stone artefacts and modified bones from Majuangou. **a**, Notch made on a flake (MJG-III). **b**, Chopper made on an angular fragment (MJG-II). **c**, Multi-platform polyhedron made on an angular fragment (MJG-III). **d**, Scraper made on a flake (MJG-III). **e**, Hammerstone; arrow indicates the main battered end. **f**, Two mammalian long-bone shaft fragments with impact notches and flake scars (arrows) typical of tool percussion damage (MJG-III). Scale bars, 1 cm.

The ages for MJG-I, MJG-II and MJG-III are considerably older than previous age estimates of Palaeolithic sites in northern China¹ and indicate that humans might have reached northeast Asia earlier than previously thought. Along with estimated ages for the sites of Gongwangling (1.15 Myr)¹⁴ and Xihoudu (1.27 Myr)¹⁵ in the southern Loess Plateau and for Xiaochangliang (1.36 Myr)¹ and Donggutuo (1.1 Myr)¹⁶ sites in the Nihewan basin, our new results imply an expansion and lengthy flourishing of human groups from northern to north-central China during the early Pleistocene.

The estimated age of 1.66 Myr for the MJG-III artefact layer pre-dates the previous oldest age of unambiguous human presence at 40° N in East Asia by about 0.3 Myr. Our findings, particularly for the MJG-III layer, document the oldest coexistence of stone tools and man-made bone modifications in East Asia, indicating possible continuity with the oldest stone tools and artificial bone modifications reported in eastern Africa^{17,18}. Archaeological evidence at MJG indicates the oldest known use of animal tissues, especially marrow processing, by early humans in Asia. The earliest archaeological level in the Nihewan basin is slightly younger than the 1.75 Myr estimated age for early humans at the Dmanisi site at 40° N latitude in western Eurasia^{2,3}. Our estimated ages also fall within the 1.66–1.51-Myr range for the earliest known human fossils in south-east Asia^{19,20}. The combined evidence suggests that, near the start of the Pleistocene, early human populations spread relatively rapidly across Asia, presumably from an African origin, and reached at least 40° N latitude. Our findings further establish that the earliest populations to reach northeast Asia were able to survive for at least 500 kyr before the mid-Pleistocene onset of high-amplitude climate oscillation^{21–23}. □

Received 19 February; accepted 8 July 2004; doi:10.1038/nature02829.

1. Zhu, R. X. *et al.* Earliest presence of humans in northeast Asia. *Nature* **413**, 413–417 (2001).
2. Gabunia, L. *et al.* Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: Taxonomy, geological setting, and age. *Science* **288**, 1019–1025 (2000).
3. Vekua, A. *et al.* A new skull of early *Homo* from Dmanisi, Georgia. *Science* **297**, 85–89 (2002).
4. Wei, Q. Banshan Paleolithic site from the lower Pleistocene in the Nihewan Basin in northern China. *Acta Anthropol. Sinica* **13**, 223–238 (1994).
5. HPICR, *Papers on Archaeology in Hebei Province* 30–45 (East, Beijing, 1998).
6. Potts, R. *Early Hominid Activities at Olduvai* (Aldine de Gruyter, New York, 1988).
7. Potts, R., Behrensmeyer, A. K. & Ditchfield, P. Paleolandscape variation and Early Pleistocene hominid activities: Members 1 and 7, Olorgesailie Formation, Kenya. *J. Hum. Evol.* **37**, 747–788 (1999).
8. Tang, Y. J., Li, Y. & Chen, W. Y. Mammalian fossils and the age of Xiaochangliang paleolithic site of Yangyuan, Hebei. *Vertebrata Palasiatica* **33**, 74–83 (1995).
9. Berggren, W. A., *et al.* in *Geochronology, Timescales, and Stratigraphic Correlation* (eds Berggren, W. A., Kent, D. V., Aubry, M. & Hardenbol, J.) 129–212 (SEPM Spec. Publ. 54, Tulsa, Oklahoma, 1995).
10. Wei, Q., *et al.* in *Evidence for Evolution—Essays in Honor of Prof. Chungchien Yong on the Hundredth Anniversary of His Birth* (ed. Tong, Y.) 193–207 (Ocean, Beijing, 1997).
11. Huang, W. P. & Fang, Q. R. *Wushan Hominid Site* 105–109 (Ocean, Beijing, 1991).
12. Qiu, Z. X. Nihewan fauna and Q/N boundary in China. *Quat. Sci.* **20**, 142–154 (2000).
13. Singer, B. S. *et al.* Dating transitionally magnetized lavas of the late Matuyama chron: Toward a new ⁴⁰Ar/³⁹Ar timescale of reversals and events. *J. Geophys. Res.* **104**, 679–693 (1999).
14. An, Z. S. & Ho, C. K. New magnetostratigraphic dates of Lantian *Homo erectus*. *Quat. Res.* **32**, 213–221 (1989).
15. Zhu, R., An, Z., Potts, R. & Hoffman, K. A. Magnetostratigraphic dating of early humans in China. *Earth Sci. Rev.* **61**, 341–359 (2003).
16. Quaternary Research Association of China, Li, H. M. & Wang, J. D. *Quaternary Geology and Environment of China* 33–38 (Ocean, Beijing, 1982).
17. Semaw, S. *et al.* 2.5-million-year-old stone tools from Gona, Ethiopia. *Nature* **385**, 333–336 (1995).
18. de Heinzelin, J. *et al.* Environment and behavior of 2.5-million-year-old Bouri hominids. *Science* **284**, 625–629 (1999).
19. Swisher, C. C. III *et al.* Age of the earliest known hominids in Java, Indonesia. *Science* **263**, 1118–1121 (1994).
20. Larick, R. *et al.* Early Pleistocene ⁴⁰Ar/³⁹Ar ages for Bapang Formation hominins, Central Java, Indonesia. *Proc. Natl Acad. Sci. USA* **98**, 4866–4871 (2001).
21. Potts, R. in *Human Roots: Africa and Asia in the Middle Pleistocene* (eds Barham, L. & Robson-Brown, K.) 5–21 (Western Academic & Specialist Press, Bristol, 2001).
22. Clark, P. U., Alley, R. B. & Pollard, D. Northern Hemisphere ice-sheet influences on global climate change. *Science* **286**, 1104–1111 (1999).
23. Tian, J., Wang, P., Cheng, X. & Li, Q. Astronomically tuned Plio-Pleistocene benthic $\delta^{18}\text{O}$ record from South China Sea and Atlantic-Pacific comparison. *Earth Planet. Sci. Lett.* **203**, 1015–1029 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. J. Enkin for providing palaeomagnetic software. This work was supported by the National Natural Science Foundation of China and Chinese Academy of Sciences. R.P. was supported by the US National Science Foundation and the Smithsonian Human Origins Program. K.A.H. also received support from the US National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.X.Z. (rxzhucn@yahoo.com and rxzhu@mail.igcas.ac.cn) or R.P. (potts.rick@nmnh.si.edu).

Modelling the recent common ancestry of all living humans

Douglas L. T. Rohde¹, Steve Olson² & Joseph T. Chang³

¹Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

²7609 Seaboard Road, Bethesda, Maryland 20817, USA

³Department of Statistics, Yale University, New Haven, Connecticut 06520, USA

If a common ancestor of all living humans is defined as an individual who is a genealogical ancestor of all present-day people, the most recent common ancestor (MRCA) for a randomly mating population would have lived in the very recent past^{1–3}. However, the random mating model ignores essential aspects of population substructure, such as the tendency of individuals to choose mates from the same social group, and the relative isolation of geographically separated groups. Here we show that recent common ancestors also emerge from two models incorporating substantial population substructure. One model, designed for simplicity and theoretical insight, yields explicit mathematical results through a probabilistic analysis. A more elaborate second model, designed to capture historical population dynamics in a more realistic way, is analysed computationally through Monte Carlo simulations. These analyses suggest that the genealogies of all living humans overlap in remarkable ways in the recent past. In particular, the MRCA of all present-day humans lived just a few thousand years ago in these models. Moreover, among all individuals living more than just a few thousand years earlier than the MRCA, each present-day human has exactly the same set of genealogical ancestors.

In investigations of the common ancestors of all living humans, much attention has focused on descent through either exclusively maternal or exclusively paternal lines, as occurs with mitochondrial DNA and most of the Y chromosome^{4,5}. But according to the more common genealogical usage of the term ‘ancestor’, ancestry encompasses all lines of descent through both males and females, so that the ancestors of an individual include all of that person’s parents, grandparents, and so on.

For a population of size n , assuming random mating (and so ignoring population substructure), probabilistic analysis² has proved that the number of generations back to the MRCA, T_n , has a distribution that is sharply concentrated around $\log_2 n$. We express this using the notation $T_n \sim \log_2 n$, meaning that the quotient $T_n/\log_2 n$ converges in probability to 1 as $n \rightarrow \infty$. In contrast, the mean time to the MRCA along exclusively matrilineal or patrilineal lines is approximately n generations⁶, and the distribution is not sharply concentrated. For example, in a panmictic population of one million people, the genealogical MRCA would have lived about 20 generations ago, or around the year AD 1400, assuming a generation time of 30 years. The MRCA along

Box 1

Graph-theoretical definitions

The length of a path in a graph, G , is the number of edges in the path. For each pair of nodes i and j in G , the distance $d(i, j)$ is defined to be the length of a shortest path joining i and j . The radius of G is

$$R = \min_{i \in G} \{ \max_{k \in G} d(i, k) \}$$

and a node i is called a centre of G if $\max_{k \in G} d(i, k) = R$. Assume $R \geq 1$; the case $R = 0$ (G has one node) was treated previously². For each centre node i , let S_i be a set of minimal size that consists of neighbours of node i and satisfies $\min \{d(j, k) : j \in \{i\} \cup S_i\} \leq R - 1$ for all $k \in G$. H_i is defined as the number of nodes in S_i , H is the minimum of H_i over all centres i , and $\Delta = 1 - \frac{1}{H}$. The diameter of G is $D = \max_{i, k \in G} d(i, k)$.

exclusively maternal lines would have lived something like 50,000 times earlier—in the order of one million generations ago.

As genealogical ancestry is traced back beyond the MRCA, a growing percentage of people in earlier generations are revealed to be common ancestors of the present-day population. Tracing further back in time, there was a threshold, let us say U_n generations ago, before which ancestry of the present-day population was an all or nothing affair. That is, each individual living at least U_n generations ago was either a common ancestor of all of today's humans or an ancestor of no human alive today. Thus, among all individuals living at least U_n generations ago, each present-day human has exactly the same set of ancestors. We refer to this point in time as the identical ancestors (IA) point. As with the MRCA point, the IA point is also quite recent in a randomly mating population: $U_n \sim 1.77 \log_2 n$ generations ago².

The major problem in applying these results to human populations is that mating is not random in the real world. Mating patterns are structured by geography, proximity, culture, language and social class. Nevertheless, even in populations with considerable internal structure, the time to the MRCA can be remarkably brief. To demonstrate this in a tractable mathematical model, consider a population of size n divided into randomly mating subpopulations that are linked by occasional migrants. The population is represented by a graph, G , with a node for each subpopulation. Edges indicate pairs of nodes that exchange a small number (for example, one pair) of migrants per generation. Let R denote the radius of G , and let Δ be a quantity ranging between 0 and 1 that depends on the structure of G (see Box 1). A probabilistic analysis (see Supplementary Information) shows that as $n \rightarrow \infty$,

$T_n \sim (R + \Delta) \log_2 n$. Furthermore, if we let D denote the diameter of the graph, then the number of generations, U_n , since the IA point satisfies $U_n \sim (D + 1.77) \log_2 n$.

Computer simulations accord with these theoretical predictions. Tables 1 and 2 give distributions of T_n and U_n for small populations of varying sizes in graphs with one node, three connected nodes, five fully connected nodes and for a ten-node graph loosely based on world geography as shown in Fig. 1. In these simulations, neighbouring subpopulations exchange one pair of migrants per generation. Each mean is calculated from 100 model runs. Although guaranteed to be accurate only for sufficiently large n , the theoretical predictions describe the simulations quite well even for models with just a few thousand individuals. Whenever n is doubled, T_n is expected to increase by $R + \Delta$, and U_n is expected to increase by $D + 1.77$. These predicted increases, which are listed in the last columns of Tables 1 and 2, agree closely with the simulation results.

To hazard a rough first guess about human recent common ancestors, we could extrapolate the results for the graph of Fig. 1 to a growing population with a final size of 250 million. When applying this model to a growing population, the fixed population size that provides the best approximation is the size at the time that the MRCA lived. We take this effective population size to be 250 million, which is approximately the global population in the year AD 1. Starting from $n = 16,000$, a population of 250 million is reached by doubling 14 times. Approximating the increases in T_n and U_n beyond the values seen in Tables 1 and 2 by their theoretical predictions for each doubling of n , we arrive at $T_n \approx 34 + 14 \times 3 = 76$ generations (about 2,300 years) and $U_n \approx 74 + 14 \times 6.77 = 169$ generations (about 5,000 years). These estimates would suggest, with the exchange of just one pair of migrants per generation between large panmictic populations of realistic size, that the MRCA appears in about the year 300 BC, and all modern individuals have identical ancestors by about 3,000 BC. Such estimates are extremely tentative, and the model contains several obvious sources of error, as it was motivated more by considerations of theoretical insight and tractability than by realism. Its main message is that substantial forms of population subdivision can still be compatible with very recent common ancestors.

The dynamics of human subpopulations are much more complex than those in the simple graph model discussed above. Although these complexities make theoretical analysis difficult, a computer model incorporating more complicated forms of population substructure and migration allows the demographic history of human populations to be simulated. The Supplementary Information contains more details on the model and computations; here we briefly outline some of the main points.

This model is based on a simplified projection of the world's

Table 1 Simulations of T_n

Graph	$n = 1,000$	$n = 2,000$	$n = 4,000$	$n = 8,000$	$n = 16,000$	$R + \Delta$
One node	10.8 (0.4)	11.8 (0.4)	12.8 (0.4)	13.9 (0.3)	14.8 (0.4)	1.00
Three fully connected nodes	14.0 (0.7)	15.6 (0.7)	17.1 (0.9)	18.9 (0.8)	20.3 (1.0)	1.50
Five fully connected nodes	14.0 (0.5)	15.8 (0.5)	17.8 (0.5)	19.6 (0.5)	21.5 (0.6)	1.75
Ten-node graph shown in Fig. 1	21.1 (1.3)	24.3 (1.5)	27.6 (1.5)	30.5 (1.5)	33.8 (1.7)	3.00

Means (standard deviations in parentheses) of T_n (the number of generations back to the MRCA) for graph-structured populations exchanging a single pair of migrants per edge per generation. The last column shows $R + \Delta$, the expected asymptotic increase in T_n per doubling of n .

Table 2 Simulations of U_n

Graph	$n = 1,000$	$n = 2,000$	$n = 4,000$	$n = 8,000$	$n = 16,000$	$D + 1.77$
One node	20.8 (1.6)	22.6 (1.5)	24.6 (1.5)	26.5 (1.6)	28.3 (1.4)	1.77
Three fully connected nodes	27.4 (1.5)	30.3 (1.4)	33.4 (1.5)	36.2 (1.7)	38.9 (1.5)	2.77
Five fully connected nodes	25.9 (1.3)	28.9 (1.4)	32.1 (1.7)	35.3 (1.5)	37.9 (1.4)	2.77
Ten-node graph shown in Fig. 1	46.3 (2.7)	53.0 (2.7)	59.8 (2.7)	66.8 (2.9)	73.6 (2.7)	6.77

Means (standard deviations in parentheses) of U_n (the number of generations back to the IA point) for graph-structured populations exchanging a single pair of migrants per edge per generation. The last column shows $D + 1.77$, the expected asymptotic increase in U_n per doubling of n .

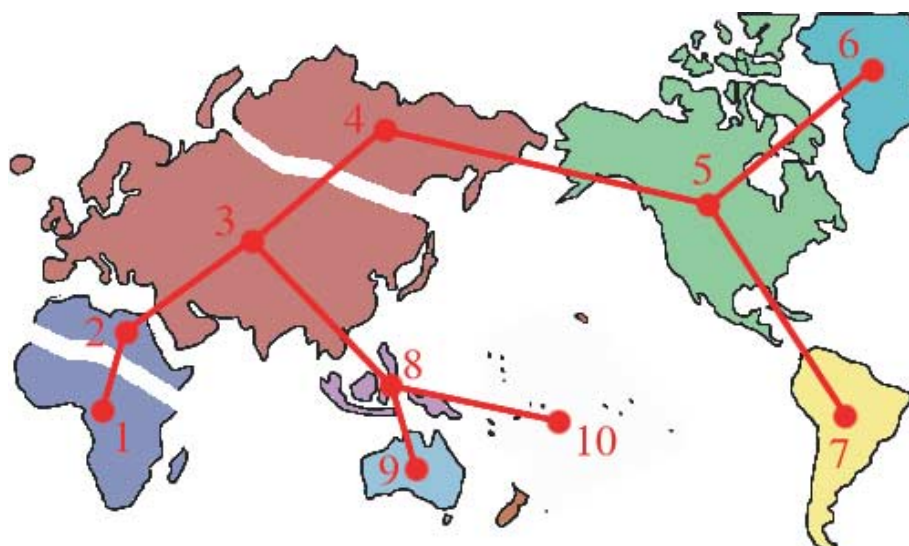


Figure 1 World map viewed as a ten-node graph. This graph has radius 3 and diameter 5.

actual inhabited land masses and has three levels of substructure: continents, 'countries' and 'towns.' Figure 2 depicts the model's geography and migration routes used before AD 1500, with the countries shown as squares and the number of towns per country differing from continent to continent. Towns and countries represent both the local geographical areas and the relevant social and ethnic groups from which most people find mates.

The model uses a simplified migration system in which each person has a single opportunity to migrate from his or her town of birth. The probabilities of leaving a town or a country are set at various levels to reflect different migration patterns. Migrants who move between towns can travel to any other town within the country. A migrant who leaves a country for another country within the same continent chooses the destination with a probability that diminishes as the inverse square of the geographical distance.

Each continent has a number of port countries from which migrants can travel to another continent. A fixed, large percentage (for example, 95% in some simulations) of the migrants through a

port come from the country in which the port is located, with the remainder drawn from other countries in the continent in proportion to their inverse squared distance. The value next to a port in Fig. 2 is its migration rate, in people per generation, and the date in parentheses indicates when the port opens, if it is more recent than the start of the simulation in 20,000 BC. When a port opens, there is usually a single generation of migration at a higher rate than the steady-state rate shown in the figure. After the year AD 1500, additional large ports, which are not shown, begin to open to simulate colonization of the Americas, Australia and elsewhere. Immediately before this, the native population of the Americas is markedly reduced to simulate the effects of European-introduced diseases⁷.

Generations overlap in this model and we explicitly simulated the lifespan and the times at which mating and reproduction events occur for each individual^{8,9}, as described in more detail in Supplementary Information. The birth rate of each continent or island was individually adjusted so that the populations match historical

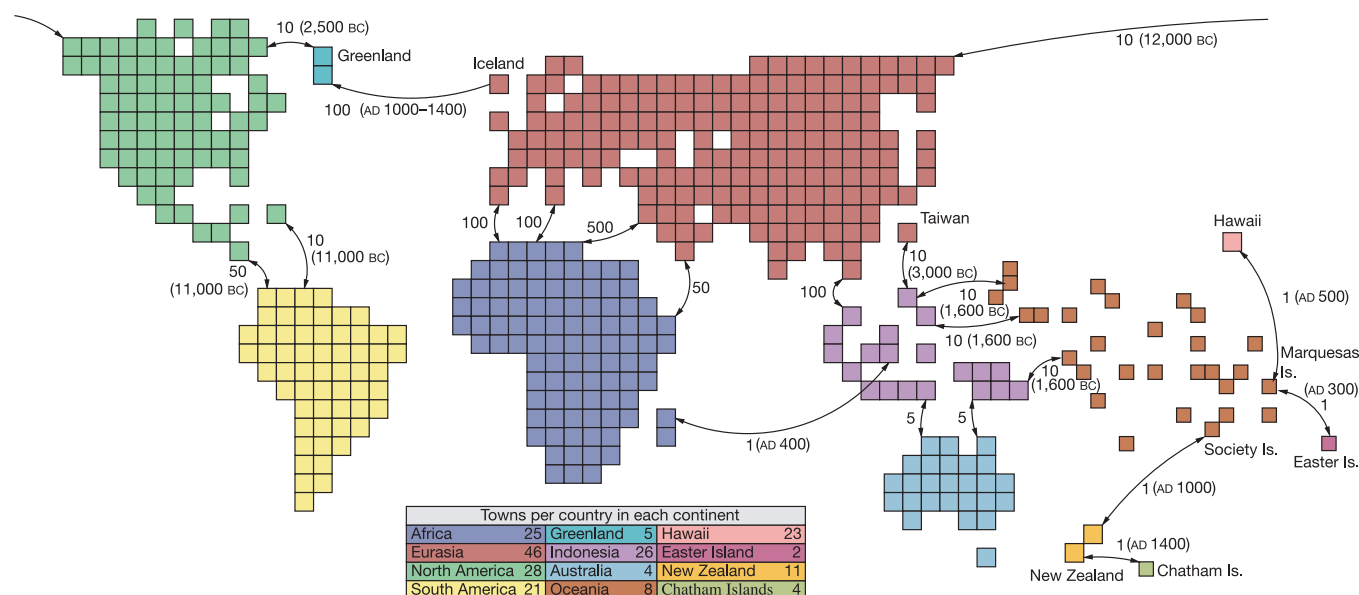


Figure 2 Geography and migration routes of the simulated model. Arrows denote ports and the adjacent numbers are their steady migration rates, in individuals per generation. If

given, the date in parentheses indicates when the port opens. Upon opening, there is usually a first-wave migration burst at a higher rate, lasting one generation.

estimates, and growth rates were higher in under-populated areas. Full-sized populations were used until the world population reached 50 million in 1,000 BC. Subsequently, birth rates were reduced to achieve a worldwide level of 55 million, carried out in such a way that sparsely populated areas were less affected. This limit was a computational necessity, but simulations show that population growth has little effect, especially if it occurs after the MRCA has died.

With 5% of individuals migrating out of their home town, 0.05% migrating out of their home country, and 95% of port users born in the country from which the port emanates, the simulations produce a mean MRCA date of 1,415 BC and a mean IA date of 5,353 BC. Interestingly, the MRCAs are nearly always found in eastern Asia. This is due to the proximity of this region to both Eurasia and either the remote Pacific islands or the Americas, allowing the MRCA's descendants to reach a few major world regions in a relatively short time.

Arguably, this simulation is far too conservative, especially given its prediction that, even in densely populated Eurasia, only 55.3 people will leave each country per generation in AD 1500. If the migration rate among towns is increased to 20%, the local port users are reduced to 80%, and the migration rates between countries and continents are scaled up by factors of 5 and 10, respectively, the mean MRCA date is as recent as AD 55 and the mean IA date is 2,158 BC. The predictions of the simple ten-node graph model sketched earlier fall somewhere between these dates and those of the more conservative computational model.

The model also can be used to calculate the percentage of ancestry that current individuals receive from different parts of the world. In generations sufficiently far removed from the present, some ancestors appear much more often than do others on any current individual's family tree, and can therefore be expected to contribute proportionately more to his or her genetic inheritance^{1,10,11}. For example, a present-day Norwegian generally owes the majority of his or her ancestry to people living in northern Europe at the IA point, and a very small portion to people living throughout the rest of the world. Furthermore, because DNA is inherited in relatively large segments from ancestors, an individual will receive little or no actual genetic inheritance from the vast majority of the ancestors living at the IA point¹².

Several factors could cause the time to the true MRCA or IA point to depart from the predictions of our model. If a group of humans were completely isolated, then no mixing could occur between that group and others, and the MRCA would have to have lived before the start of the isolation. A more recent MRCA would not arise until the groups were once again well integrated. In the case of Tasmania, which may have been completely isolated from mainland Australia between the flooding of the Bass Strait, 9,000–12,000 years ago, and the European colonization of the island, starting in 1803 (ref. 13), the IA date for all living humans must fall before the start of isolation. However, the MRCA date would be unaffected, because today there are no remaining native Tasmanians without some European or mainland Australian ancestry.

No large group is known to have maintained complete reproductive isolation for extended periods. The populations on either side of the Bering Strait appear to have exchanged mates throughout the period documented in the archaeological record¹⁴. Religious isolates such as the Samaritans occasionally have absorbed migrants from outside the group¹⁵. Even populations on isolated Pacific islands have experienced occasional infusions of newcomers¹⁶. Even if rates of migration between some adjoining populations are very low, the time to the MRCA tends not to change substantially. For example, with a migration rate across the Bering Strait of just one person in each direction every ten generations, rather than the ten per generation in the more conservative simulation described earlier, T_n only increases from 3,415 years to 3,668 years.

Conversely, other factors could reduce the time to the MRCA

from that predicted by the model. Examples of such factors include the existence of more diverse intercontinental migration routes, the large-scale movement and mixing of populations documented in the historical record¹⁷, marked individual differences in fertility¹⁸, and the population increase of the past two millennia, which would result in more migrants.

Actual migration rates among populations are very poorly known and undoubtedly have varied considerably in different times and places. Studies of hunter-gatherer groups and subsistence agricultural communities have found that anywhere from 1% (ref. 19) to as much as 30% (ref. 20) of mates are from outside the group. The tendency of most human groups to marry out with surrounding groups, at least to a limited extent, links networks of ancestry within specific regions (see <http://www.compapp.dcu.ie/~humphrys/FamTree/Royal/Famous.descents.html>).

Given the remaining uncertainties about migration rates and real-world mating patterns, the date of the MRCA for everyone living today cannot be identified with great precision. Nevertheless, our results suggest that the most recent common ancestor for the world's current population lived in the relatively recent past—perhaps within the last few thousand years. And a few thousand years before that, although we have received genetic material in markedly different proportions from the people alive at the time, the ancestors of everyone on the Earth today were exactly the same.

Further work is needed to determine the effect of this common ancestry on patterns of genetic variation in structured populations^{21–24}. But to the extent that ancestry is considered in genealogical rather than genetic terms, our findings suggest a remarkable proposition: no matter the languages we speak or the colour of our skin, we share ancestors who planted rice on the banks of the Yangtze, who first domesticated horses on the steppes of the Ukraine, who hunted giant sloths in the forests of North and South America, and who laboured to build the Great Pyramid of Khufu. □

Received 30 December 2003; accepted 14 July 2004; doi:10.1038/nature02842.

1. Wachter, K. W. in *Genealogical Demography* (eds Dyke, B. & Morrill, W. T.) 85–93 (Academic, New York, 1980).
2. Chang, J. T. Recent common ancestors of all present-day individuals. *Adv. Appl. Probab.* **31**, 1002–1026, 1027–1038 (1999).
3. Derrida, B., Manrubia, S. C. & Zanette, D. H. On the genealogy of a population of biparental individuals. *J. Theor. Biol.* **203**, 303–315 (2000).
4. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllenstein, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
5. Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. & Feldman, M. W. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc. Natl Acad. Sci. USA* **97**, 7360–7365 (2000).
6. Hudson, R. R. in *Oxford Surveys of Evolutionary Biology* (eds Harvey, P. H. & Partridge, L.) 1–44 (Oxford Univ. Press, New York, 1990).
7. Stannard, D. E. *American Holocaust: Columbus and the Conquest of the New World* (Oxford Univ. Press, New York, 1992).
8. US National Office of Vital Statistics, *Death Rates by Age, Race, and Sex, United States, 1900–1953, Vital Statistics—Special Reports Vol. 43* (US Government Printing Office, Washington DC, 1956).
9. Pletcher, S. D. Model fitting and hypothesis testing for age-specific mortality data. *J. Evol. Biol.* **12**, 430–439 (1999).
10. Ohno, S. The Malthusian parameter of ascents: What prevents the exponential increase of one's ancestors? *Proc. Natl Acad. Sci. USA* **93**, 15276–15278 (1996).
11. Derrida, B., Manrubia, S. C. & Zanette, D. H. Distribution of repetitions of ancestors in genealogical trees. *Physica A* **281**, 1–16 (2000).
12. Wiuf, C. & Hein, J. On the number of ancestors to a DNA sequence. *Genetics* **147**, 1459–1468 (1997).
13. Jones, R. Tasmanian archaeology: Establishing the sequences. *Ann. Rev. Anthropol.* **24**, 423–446 (1995).
14. Fitzhugh, W. W. & Chaussonnet, V. (eds) *Crossroads of Continents: Cultures of Siberia and Alaska* (Smithsonian Institution Press, Washington DC, 1988).
15. Bonnè-Tamir, B. et al. Maternal and paternal lineages of the Samaritan isolate: Mutation rates and time to most recent common male ancestor. *Ann. Hum. Genet.* **67**, 153–164 (2003).
16. Morton, N. E., Harris, D. E., Yee, S. & Lew, R. Pingelap and Mokil atolls: Migration. *Am. J. Hum. Genet.* **23**, 339–349 (1971).
17. Hoerder, D. *Cultures in Contact: World Migrations in the Second Millennium* (Duke Univ. Press, Durham, North Carolina, 2002).
18. Zerjal, T. et al. The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72**, 717–721 (2003).
19. Weiss, K. M. & Maruyama, T. Archeology, population genetics and studies of human racial ancestry. *Am. J. Phys. Anthropol.* **44**, 31–50 (1976).
20. Ward, R. H. & Neel, J. V. Gene frequencies and microdifferentiation among the Makiritare indians. IV. A comparison of a genetic network with ethnohistory and migration matrices; a new index of genetic isolation. *Am. J. Hum. Genet.* **22**, 538–561 (1970).

21. Jorde, L. B. in *Current Developments in Anthropological Genetics* (eds Mielke, J. H. & Crawford, M. H.) 135–208 (Plenum, New York, 1980).
22. Notohara, M. The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.* **29**, 59–75 (1990).
23. Wilkinson-Herbots, H. M. Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37**, 535–585 (1998).
24. Hey, J. & Machado, C. A. The study of structured populations—new hope for a difficult and divided science. *Nature Rev. Genet.* **4**, 535–543 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The research of D.L.T.R. was supported by the National Institutes of Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.L.T.R. (dr@tedlab.mit.edu).

Phenotypic consequences of 1,000 generations of selection at elevated CO₂ in a green alga

Sinéad Collins & Graham Bell

Biology Department, McGill University, Montreal, Quebec H3A 1B1, Canada

Estimates of the effect of increasing atmospheric CO₂ concentrations on future global plant production rely on the physiological response of individual plants or plant communities when exposed to high CO₂ (refs 1–6). Plant populations may adapt to the changing atmosphere, however, such that the evolved plant communities of the next century are likely to be genetically different from contemporary communities^{7–12}. The properties of these future communities are unknown, introducing a bias of unknown sign and magnitude into projections of global carbon pool dynamics. Here we report a long-term selection experiment to investigate the phenotypic consequences of selection for growth at elevated CO₂ concentrations. After about 1,000 generations, selection lines of the unicellular green alga *Chlamydomonas* failed to evolve specific adaptation to a CO₂ concentration of 1,050 parts per million. Some lines, however, evolved a syndrome involving high rates of photosynthesis and respiration, combined with higher chlorophyll content and reduced cell size. These lines also grew poorly at ambient concentrations of CO₂. We tentatively attribute this outcome to the accumulation of conditionally neutral mutations in genes affecting the carbon concentration mechanism.

Plant growth depends on CO₂ concentration^{1,2}, which is expected to rise from current levels of about 400 parts per million (p.p.m.) to between 700 and 1,000 p.p.m. during the next century³. In response, global plant productivity in forests⁴, grasslands⁵, agroecosystems⁶ and other ecosystems is expected to increase. Projections of future net primary productivity are complicated by synchronous changes in temperature and other factors, but most models predict increases in the land–atmosphere and ocean–atmosphere fluxes from current values of >–2 petagrams (Pg) C per year to about –5 Pg C per year³. This process is likely to be complicated by shifts in the species composition of plant communities⁷, and more fundamentally by evolutionary changes within plant populations. In the very long term, this may involve the extinction of some groups and the radiation of others⁸, but within a few hundred generations most plant populations may adapt to the increased supply of inorganic carbon. Selection experiments with plants have demonstrated a variety of

responses^{9–12}, but have been limited to fewer than ten generations. The long-term response to selection and the properties of populations adapted to elevated CO₂ remain unknown, and constitute an important limit on our ability to predict future plant productivity.

We used a microbial model system in which large population size and short generation time make it possible to evaluate evolutionary change caused by the spread of novel mutations over hundreds of generations. *Chlamydomonas reinhardtii* is a unicellular green alga that has been extensively used to study the physiology and genetics of photosynthesis¹³. It possesses a carbon-concentrating mechanism (CCM), which increases the concentration of CO₂ near the active site of ribulose 1,5-bisphosphate carboxylase–oxygenase (Rubisco), in common with most other eukaryotic microalgae that have been studied¹⁴. We set up ten isogenic selection lines from each of two ancestral genotypes, half being grown at ambient CO₂ (ambient lines) and half at a concentration that increased from ambient to 1,050 p.p.m. over about 600 generations and was then maintained at this level for a further 400 generations (high lines). At least 10⁵ cells per line were transferred for 125 transfers in a buffered, nutrient-rich medium. The history of these lines thus emulates the conditions that photosynthetic organisms are likely to experience during the next century or so, with respect to CO₂ levels alone.

The physiological effect of elevated CO₂ concentration is expected to be an increase in photosynthesis, causing an increase in growth. Net photosynthesis in the ambient lines increased by about 30% when they were grown at high CO₂ (Fig. 1a). The ambient lines diverged through time so that by the end of the experiment they varied significantly in the rate of photosynthesis (one-way analysis of variance (ANOVA): $F_{9,18} = 9.0$, $P < 0.001$) when grown at ambient CO₂ concentrations. The high lines had normal rates of photosynthesis at ambient CO₂, which increased by more than 50% as an average over all lines at high CO₂. However, this effect was very inconsistent: one group of high lines had low rates whereas a second group had very high rates of photosynthesis at high CO₂ concentration (Fig. 1a). This distinction was not related to the identity of the ancestor, and represented significantly more divergence in photosynthetic rates than was seen in the ambient lines ($F_{1,16} = 10.5$, $P = 0.005$).

The growth rate of cultures grown at elevated CO₂ was correlated with their photosynthetic rate among the ambient lines, but not among the high lines (Fig. 1b). The physiological effect of CO₂ on photosynthesis was reflected by growth in pure culture, where the maximal rate of increase (Fig. 1c) and the limiting density (Fig. 1d) of both the ambient and the high lines are enhanced substantially by high CO₂. However, there was no indication of a parallel evolutionary response: by the end of the selection experiment, the high lines had not become specifically adapted to growth at high CO₂; their growth at high CO₂ being no greater than, and perhaps even less than, the growth of the ambient lines. There was nevertheless an indirect response: the growth of some high lines was markedly impaired at ambient CO₂ concentrations where two of the lines could scarcely be propagated. This result was supported by the outcome of competition assays in which the selection lines were mixed with standard genetically marked strains and the change in frequency during growth in culture recorded (Table 1). The high lines had considerably lower competitive ability at ambient CO₂, where three of them (including the two with strongly reduced growth in pure culture) were such weak competitors that they were consistently eliminated by the tester strains within 10–15 generations. They were, however, no more successful than the ambient lines at high CO₂. In short, 1,000 generations of selection at high CO₂ concentrations had caused no increase in growth at high CO₂, whereas growth at ambient CO₂ was often considerably reduced.

Photosynthesis is linearly related to respiration in the dark among lines at ambient CO₂; this relationship is the same for ambient and high lines (Fig. 2a). It has been shown in *Chlamydomonas* that post-illumination rates of O₂ consumption provide a

good estimate of the rates of respiratory O_2 consumption during the preceding light period¹⁵. The same correlation is expressed at high CO_2 , but respiration rates are on average greater for the high lines, because some high lines have extremely high respiration rates, whereas others have normal rates (Fig. 2b). Two of the three high lines that had high photosynthetic rates also had very high respiration rates. This will reduce the effectiveness of photosynthesis, yet it cannot alone account for the lack of increase in growth, because our measurements of net photosynthesis include photorespiration. In these cases, another process, such as increased leakage of fixed carbon from cells, must also be involved. The evolutionary response is in the opposite sense to the physiological response, which for elevated CO_2 concentrations is to induce lower rates of dark respiration in C_3 land plants².

Both chlorophyll content and cell size responded to selection at elevated CO_2 concentrations. The physiological response to increased CO_2 is an increase in chlorophyll *a* content, seen in both the ambient and high lines. In the ambient lines, the average increase in chlorophyll content per cell is about 28%. The high lines, however, show the same inconsistent effect as with rates of photosynthesis (Fig. 3a): those lines with very high photosynthetic rates at high CO_2 also have very large increases in chlorophyll content at high CO_2 . The physiological effect of high CO_2 is a marked increase in cell volume, both in ambient and in high lines (Fig. 3b). The high lines have smaller cells than the ambient lines, regardless of the

Table 1 Competitive fitness

Selection environment	Assay environment	Competitive fitness ($\pm$ s.e.)	Extinct
High CO_2	High CO_2	-0.682 ± 0.306	0
High CO_2	Ambient CO_2	-1.231 ± 0.325	3
Ambient CO_2	High CO_2	-0.292 ± 0.325	1
Ambient CO_2	Ambient CO_2	-0.386 ± 0.290	0

Fitness of all lines was measured against a common marked strain. Selection and marked lines were inoculated in approximately equal numbers at the beginning of the assay. Fitness is calculated such that a fitness of 0 indicates fitness equal to the marker. There is a marginally significant effect of selection on competitive fitness (ANOVA: $F = 3.92$, $P = 0.056$). When a selection line was completely out-competed by the marked strain by the first time point, it was assumed to be present at the limit of detection (1/200); the number of lines becoming 'extinct' is recorded in the final column.

concentration of CO_2 . The effect is considerable, amounting to an average reduction of 22% in volume. The evolutionary response is thus of comparable magnitude but opposite in sign to the physiological response. The evolution of reaction norms in this way, so as to mitigate unfavourable physiological responses to extreme environments, has often been observed in long-term selection experiments¹⁶.

In addition to a markedly reduced ability to grow at ambient concentrations of CO_2 , the high lines also had a lower limiting density at high CO_2 , suggestive of either a lowered affinity for CO_2 , or a higher per-cell requirement for inorganic carbon, attributable to higher internal organic carbon content, an increase in respiration

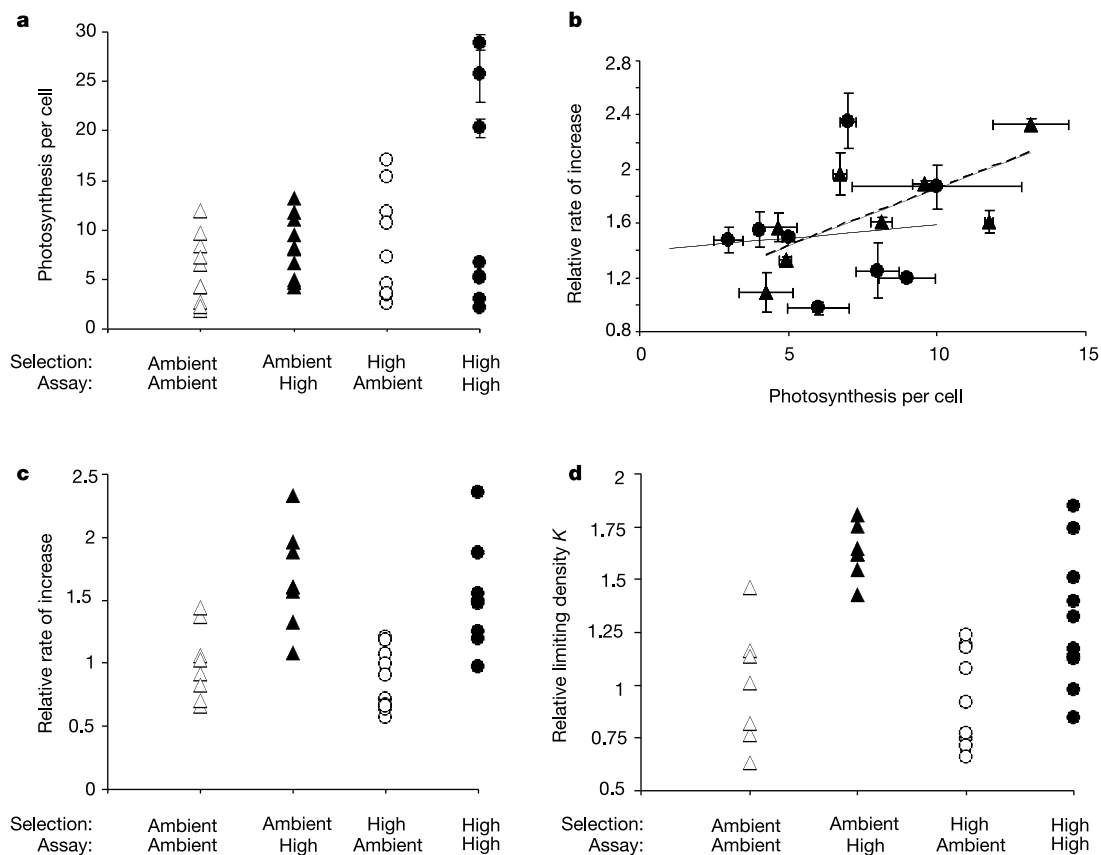


Figure 1 Response to selection at elevated CO_2 concentrations. Symbols designate conditions of selection (ambient, triangle; high, circle) and assay (ambient, open; high, filled). **a**, Photosynthetic rates measured as O_2 evolution per cell. Points are means based on two independent replicates; average s.e.m. 0.61 (range 0.02–2.85). Error bars (± 1 s.e.) are shown explicitly for the three selection lines with exceptionally high rates of photosynthesis at elevated CO_2 . **b**, Relationship between growth rate and photosynthesis at high CO_2 for ambient lines (upper regression, $P = 0.038$) and high lines (lower regression, not significant). Error bars are ± 1 s.e. **c**, Pure culture growth rates. All values

are calculated relative to the average growth rate of ambient lines growing at ambient CO_2 concentrations. Lines show increased growth rates at high CO_2 ($F = 33.6$, $P < 0.0001$). Points are means based on three independent replicates; average s.e.m. 0.096 (range 0.01–0.51). **d**, Limiting densities. All values are calculated relative to ambient lines growing at ambient CO_2 . High lines have significantly lowered carrying capacities than do ambient lines (effect of selection: $F = 5.1$, $P = 0.03$; effect of assay: $F = 32.5$, $P < 0.0001$). Points are means based on three independent replicates; average s.e.m. 0.116 (range 0.01–0.25).

or increased loss of carbon from cells. Under ambient CO_2 conditions, *Chlamydomonas* and other microalgae concentrate inorganic carbon through an energy-requiring process that keeps Rubisco saturated with CO_2 ^{17–19}. When the external concentration of CO_2 is increased, mutations in downregulated or unexpressed CCM genes might be neutral in the high- CO_2 environment, despite being deleterious in the ancestral environment. If the CCM were compromised in some way, the evolved lines would show a decreased affinity for inorganic carbon, resulting in a decreased limiting density. A lowered affinity for inorganic carbon without a necessary decrease in steady-state photosynthesis at high CO_2 is seen in some *Chlamydomonas* high- CO_2 -requiring mutants where components of the CCM are inactivated^{20–22}. We tentatively attribute the outcome of selection in our experiment to the accumulation of conditionally neutral mutations that are deleterious when expressed in more stringent conditions. This class of mutation has previously been shown to explain antagonistic indirect responses to selection in microbial selection experiments^{23,24}.

The main result of our experiment is that we were unable to demonstrate specific adaptation to high CO_2 concentrations, because after 1,000 generations the high selection lines neither grew faster nor had a higher limiting density than did the ambient lines in the high CO_2 environment. Instead, the physiological response of all traits measured was attenuated or reversed in at least some of the high lines by the end of the selection experiment. This suggests that projecting future change on the basis of current

physiological responses may be misleading, and should wherever possible be attempted in conjunction with an empirical knowledge of evolutionary responses. However, a more fundamental obstacle to precise forecasting is the uncertainty of the evolutionary response, which is a general feature of selection experiments^{25–27}. We observed the evolution of two distinct syndromes with respect to carbon metabolism, defined by the ability to respond physiologically to changes in CO_2 . One group of lines showed no change in chlorophyll *a* content, photosynthesis or respiration rates, and when growing at high CO_2 seemed to mimic control cells growing at ambient concentrations of CO_2 . The second group of selection lines was more variable in its responses, but all lines within the group had elevated rates of photosynthesis when grown at high CO_2 , although they could not channel the fixed carbon into growth. Two of the three evolved lines with elevated photosynthetic rates also had elevated respiration rates when grown at high CO_2 concentrations.

The outcome of our experiment suggests that over the next century many phytoplankton communities will evolve less efficient CCMs through the accumulation of conditionally neutral mutations, and will come to consist of smaller cells with broader ranges in photosynthesis and respiration rates than is currently seen. This would affect global processes by changing the rate of carbon turnover in aquatic and perhaps in terrestrial systems. The experimental system, however, is far from perfect in simulating the

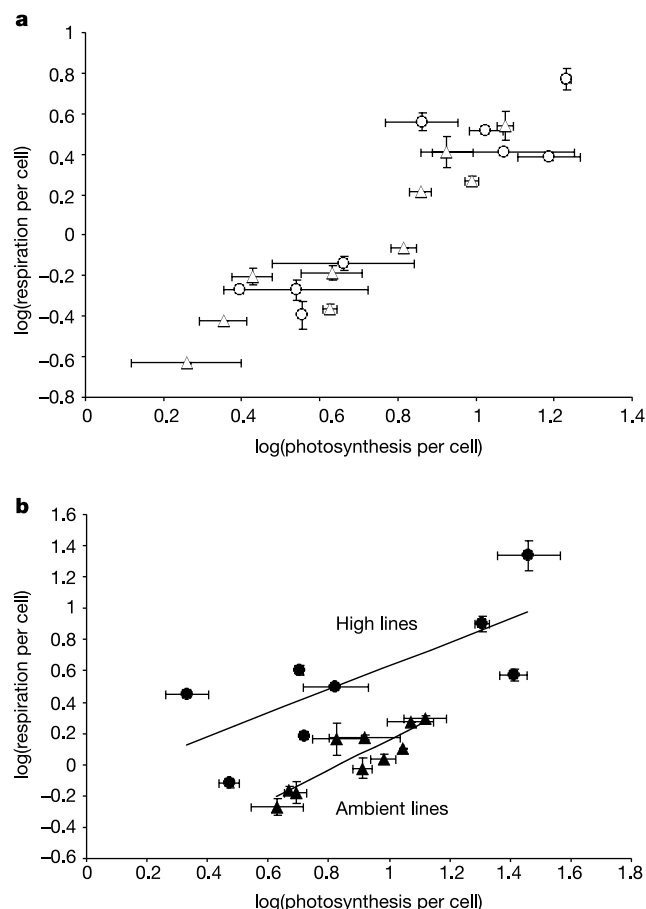


Figure 2 Relationship between photosynthesis and respiration rates at ambient (a) and high (b) CO_2 concentrations. Photosynthesis and respiration rates are measured as O_2 evolved $\text{s}^{-1} \text{ cell}^{-1}$ and O_2 consumed $\text{s}^{-1} \text{ cell}^{-1}$, respectively. Two high lines, which failed to grow at ambient CO_2 , were excluded. Each point represents independent duplicate measurements of a single line; error bars are ± 1 s.e. Circles, high lines; triangles, ambient lines.

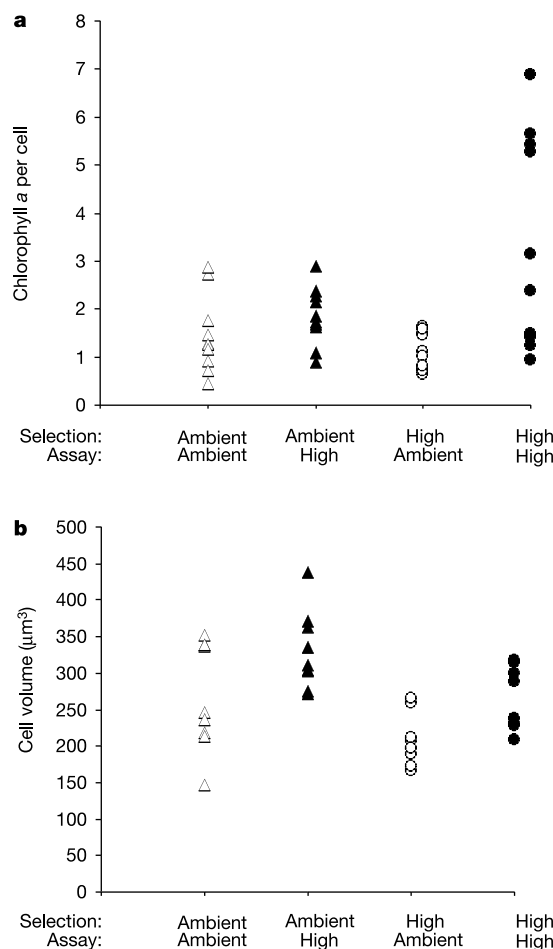


Figure 3 Correlated responses to selection at elevated CO_2 . **a**, Chlorophyll *a* content per cell. Points are means of two measurements; average s.e.m. 0.116 (range 0.01–0.92). **b**, Cell volume. Each point represents average cell volume of a single replicate line. Averages were calculated from measurements of 200 cells (high lines) or 170 cells (ambient lines); average s.e.m. 1.53 (range 0.79–2.68). Two lines where single cells could not be accurately measured because of clumping were excluded.

phytoplankton communities of oligotrophic ocean systems, and still less so for terrestrial plants. We have described phenotypes likely to result from changes in CO₂ concentrations alone; how changes in other variables, such as temperature, pH and nutrients modify these phenotypes remain to be seen. Selection experiments in more realistic systems will be necessary to validate the evolutionary response to global atmospheric change. □

Methods

Selection experiment

Ten replicate lines were founded from a single clone of *C. reinhardtii* M566B (laboratory isolate), and ten replicate lines were founded from a single clone of CC2344 (Chlamydomonas Genetics Center, Duke University). Five replicates from each clone were grown in an increasing CO₂ environment and five replicates from each clone were grown in an ambient CO₂ environment. The ambient CO₂ environment consisted of flasks being bubbled with air containing 430 p.p.m. CO₂ for the entire experiment. Lines in the high CO₂ treatment were initially grown in flasks being bubbled with air containing 430 p.p.m. CO₂, and CO₂ levels were raised steadily to 1,050 p.p.m. over the first 600 generations of the experiment. These lines were then grown at 1,050 p.p.m. CO₂ for a further 400 generations. Lines were propagated by batch culture grown in bubbled flasks containing 300 ml of Suoka high salt medium²⁸ (HSM) in a phytotron chamber under constant light (800 ± 20 µmol m⁻² s⁻¹) at 25 °C. One millilitre (about 10⁵ cells) was transferred every 3–4 days for approximately 1,000 generations for each replicate line.

Pure culture growth rates and limiting densities

Pure culture growth rates were measured in 384-well plates containing 90 µl HSM per well. Cultures were first acclimated (3–6 days), then diluted and transferred to assay plates. For the two evolved lines that often failed to grow at ambient CO₂ concentrations, several extra acclimation cultures were inoculated, and the surviving cultures were used for growth assays. The plates were grown in the same phytotron chamber as above at either 430 p.p.m. or 1,050 p.p.m. CO₂. Absorbance of each culture was measured every 24 h. Limiting densities were calculated from the maximum absorbance maintained by a culture. Values given are means of three independent replicates.

Competitive fitness assay

Competitive fitness was measured by inoculating 300 ml of HSM with equal volumes (approximately equal numbers) of acclimated selection line and a marked strain CC48 arg⁻ (from Chlamydomonas Genetics Center, Duke University). The flasks were grown in the same phytotron chamber as above, bubbled with either 430 p.p.m. or 1,050 p.p.m. CO₂. The cultures were sampled every 3 days and plated on HSM plus arginine plates. After colony growth, the plates were counted and then replica-plated onto HSM-only plates. The relative frequencies of marker and selection lines were calculated by difference. Dead (arginine-requiring) colonies were usually visible on the HSM-only plates. In cases where the selection lines were absent on plates, they were assumed to be present just below the detection limit of the assay, and were entered into the analysis as having a frequency of 0.005. Values given are means from three independent replicates.

Photosynthesis and respiration assays

Photosynthetic oxygen evolution and respiration (oxygen uptake in the dark) were measured in a Clark-type oxygen electrode illuminated at 800 µmol m⁻² s⁻¹. Cultures were depleted of oxygen by bubbling with N₂/CO₂ at either 430 p.p.m. or 1,050 p.p.m. CO₂. Net photosynthesis (oxygen electrode output from illuminated cells) was used for analysis. Respiration was calculated from oxygen uptake in the dark immediately after a light period. Values given are means of two independent replicate measurements. Chlorophyll was determined by acetone extraction²⁸. Values given are means of two replicate measurements from the same culture.

Cell measurements

Cells from acclimated liquid cultures were fixed and 200 (high lines) or 170 (ambient lines) cells measured under a microscope. Cell volume was calculated based on the shape of cells²⁹.

Received 26 May; accepted 17 August 2004; doi:10.1038/nature02945.

- Urban, O. Physiological impacts of elevated CO₂ concentration ranging from molecular to whole plant responses. *Photosynthetica* **41**, 9–20 (2003).
- Griffin, K. L. *et al.* Plant growth in elevated CO₂ alters mitochondrial number and chloroplast fine structure. *Proc. Natl Acad. Sci. USA* **98**, 2473–2478 (2001).
- Watson, R., Houghton, J. & Yihui, D. (eds) *Climate Change 2001: The Scientific Basis* (Intergovernmental Panel on Climate Change, Geneva, 2001).
- Saxe, H., Ellsworth, D. S. & Heath, J. Tree and forest functioning in an enriched CO₂ atmosphere. *New Phytol.* **139**, 395–436 (1998).
- Lee, T. D., Tjoelker, M. G., Ellsworth, D. S. & Reich, P. B. Leaf gas exchange responses of 13 prairie grassland species to elevated CO₂ and increased nitrogen supply. *New Phytol.* **150**, 405–418 (2001).
- Fuhrer, J. Agroecosystem responses to combinations of elevated CO₂, ozone and global climate change. *Agric. Ecosyst. Environ.* **97**, 1–20 (2003).
- Luescher, A., Hendrey, G. R. & Noesburger, J. Long-term responsiveness to free air CO₂ enrichment of functional types, species and genotypes of plants from fertile permanent grassland. *Oecologia* **113**, 37–45 (1998).
- Tilman, D. & Lehman, C. Human-caused environmental change: Impacts on plant diversity and evolution. *Proc. Natl Acad. Sci. USA* **98**, 5433–5440 (2001).

- Bazzaz, F. A., Jasienki, M., Thomas, S. C. & Wayne, P. Microevolutionary responses in experimental populations of plants to CO₂-enriched environments. *Proc. Natl Acad. Sci. USA* **92**, 8161–8165 (1995).
- Ward, J. K., Antonovics, J., Thomas, R. B. & Strain, B. R. Is atmospheric CO₂ a selective agent on model C3 annuals? *Oecologia* **123**, 330–341 (2000).
- Andalo, C., Goldringer, I. & Godelle, B. Inter- and intragenotypic competition under elevated carbon dioxide in *Arabidopsis thaliana*. *Ecology* **82**, 157–164 (2001).
- Potvin, C. & Toussein, D. Evolutionary consequences of simulated global change: Genetic versus plastic responses. *Oecologia* **108**, 683–693 (1997).
- Raven, J. A. & Girard-Bascou, J. Algal model systems and the elucidation of photosynthetic metabolism. *J. Phycol.* **37**, 943–950 (2001).
- Colman, B., Huertas, I. E., Bhatti, S. & Dason, J. S. The diversity of inorganic carbon acquisition mechanisms in eukaryotic microalgae. *Funct. Plant Biol.* **29**, 261–270 (2002).
- Falconer, D. S. Selection in different environments: effects on environmental sensitivity (reaction norm) and on mean performance. *Genet. Res.* **56**, 57–70 (1990).
- Sültemeyer, D. Carbonic anhydrase in eukaryotic algae: characterization, regulation, and possible function during photosynthesis. *Can. J. Bot.* **76**, 962–972 (1998).
- Badger, M. R. *et al.* The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO₂ concentrating mechanisms in algae. *Can. J. Bot.* **76**, 1052–1071 (1998).
- Badger, M. R. & Spalding, M. H. *Photosynthesis: Physiology and Metabolism* 369–397 (Kluwer, New York, 2000).
- Thyssen, C., Hermes, M. & Sültemeyer, D. Isolation and characterization of *Chlamydomonas reinhardtii* mutants with an impaired CO₂-concentrating mechanism. *Planta* **217**, 102–112 (2003).
- Suzuki, K. & Spalding, M. H. Adaptation of *Chlamydomonas reinhardtii* high-CO₂-requiring mutants to limiting CO₂. *Plant Physiol.* **90**, 1195–1200 (1988).
- Spalding, M. H., Van, K., Wang, Y. & Nakamura, Y. Acclimation of *Chlamydomonas* to changing carbon availability. *Funct. Plant Biol.* **29**, 221–230 (2002).
- Macleay, R. C. & Bell, G. Experimental adaptive radiation in *Pseudomonas*. *Am. Nat.* **160**, 569–581 (2002).
- Turner, P. E. & Elena, S. F. Cost of host radiation in an RNA virus. *Genetics* **156**, 1465–1470 (2000).
- Travisano, M. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli*. IV. Targets of selection and the specificity of adaptation. *Genetics* **143**, 15–26 (1996).
- Notley-McRobb, L. & Ferenci, T. Adaptive *mgl*-regulatory mutations and genetic diversity evolving in glucose-limited *Escherichia coli* populations. *Environ. Microbiol.* **1**, 33–43 (1999).
- Yedig, G. & Bell, G. Macroevolution simulated with autonomously replicating computer programmes. *Nature* **420**, 810–812 (2002).
- Xue, X. *et al.* Interactions between photosynthesis and respiration in the green alga *Chlamydomonas reinhardtii*. *Plant Physiol.* **112**, 1005–1014 (1996).
- Harris, E. H. *The Chlamydomonas Sourcebook: A Comprehensive Guide to Biology and Laboratory Use* (Academic, San Diego, California, 1989).
- Hillebrand, H., Dürselen, C.-D., Kirschtel, D., Pollinger, U. & Zohary, T. Biovolume calculation for pelagic and benthic microalgae. *J. Phycol.* **35**, 403–424 (1999).

Acknowledgements This research was supported by a Discovery Grant from the Natural Sciences and Engineering Council of Canada to G.B.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.B. (graham.bell@mcgill.ca).

Pack-MULE transposable elements mediate gene evolution in plants

Ning Jiang^{1*†}, Zhirong Bao^{2*†}, Xiaoyu Zhang^{1†}, Sean R. Eddy² & Susan R. Wessler¹

¹Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA
²Howard Hughes Medical Institute and Department of Genetics, Washington University, St Louis, Missouri 63108, USA

* These authors contributed equally to this study

† Present addresses: Department of Horticulture, Michigan State University, East Lansing, Michigan 48824, USA (N.J.); Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA (Z.B.); Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, California 90095, USA (X.Z.)

Mutator-like transposable elements (MULEs) are found in many eukaryotic genomes and are especially prevalent in higher plants^{1–3}. In maize, rice and *Arabidopsis* a few MULEs were shown to carry fragments of cellular genes^{4–6}. These chimaeric elements are called Pack-MULEs in this study. The abundance of MULEs in rice and the availability of most of the genome sequence permitted a systematic analysis of the prevalence and

nature of Pack-MULEs in an entire genome. Here we report that there are over 3,000 Pack-MULEs in rice containing fragments derived from more than 1,000 cellular genes. Pack-MULEs frequently contain fragments from multiple chromosomal loci that are fused to form new open reading frames, some of which are expressed as chimaeric transcripts. About 5% of the Pack-MULEs are represented in collections of complementary DNA. Functional analysis of amino acid sequences and proteomic data indicate that some captured gene fragments might be functional. Comparison of the cellular genes and Pack-MULE counterparts indicates that fragments of genomic DNA have been captured, rearranged and amplified over millions of years. Given the abundance of Pack-MULEs in rice and the widespread occurrence of MULEs in all characterized plant genomes, gene fragment acquisition by Pack-MULEs might represent an important new mechanism for the evolution of genes in higher plants.

MULEs are DNA transposons that can be classified as either autonomous (transposase-encoding) or non-autonomous (not encoding but requiring transposase). Unlike non-autonomous DNA elements from most transposon families, non-autonomous MULEs are usually not deletion derivatives of autonomous MULEs. Instead, a variety of sequences are found between MULE terminal inverted repeats (TIRs) including fragments from host genes. This phenomenon was first reported for the maize *Mu1* element, which contains part of a gene of unknown function called *MRS-A*^{4,7}. Recently a few *Arabidopsis* and rice MULEs harbouring fragments of host genes were reported^{5,6}. Such MULEs are here referred to as Pack-MULEs.

To assess the impact of Pack-MULEs on gene and genome evolution it is necessary first to evaluate the abundance of these elements. To determine the prevalence of Pack-MULEs on a genome-wide basis, we turned to rice with its nearly complete genome sequence^{8–10}. Using a computational approach (see Methods and Supplementary Methods for details), 266 families of MULE-related sequences were identified and used to search 440 megabases (Mb) of Nipponbare sequence, yielding about 67,000 non-redundant hits. These hits were used to identify Pack-MULEs by employing the following criteria: first, TIRs separated by less than 5 kilobases (kb) and internal sequences should not contain another transposable element (TE); second, TIRs should belong to the same MULE family (as identified by RECON, software developed for *de novo* repeat identification¹¹); third, TIRs should be in inverted orientation with terminal sequences outwards (as in previously described MULEs); last, the sequence between TIRs must be highly similar (BLASTX $E < 10^{-9}$) to non-transposase and non-hypothetical proteins in GenBank and in the *indica* rice genome (see Methods for details)¹⁰.

With these criteria, 1,380 Pack-MULEs were detected. To test whether the search identified genuine Pack-MULEs or whether most of the 1,380 sequences were fortuitous associations of independent MULE TIRs reflecting the high local density of MULEs (150 hits per Mb) in the genome, two control experiments were performed. First, if the TIRs were generated by independent events, their orientation should be uncorrelated. Thus, the same search but with TIRs pointed in the 'wrong' direction (inwards instead of outwards) should produce a similar number of Pack-MULEs. However, this search yielded only 40 hits, compared with the original 1,380. Second, a random sample of 100 elements and their flanking sequences was screened manually for target site duplications (TSDs; see Methods) that are generated on insertion. Because MULEs have no significant target sequence preference, most insertions should have a unique TSD^{5,6,12–14}. Thus, a recognizable TSD flanking the two TIRs indicates that the TIRs belong to a single element, rather than independent MULE ends that fortuitously flank a gene. Seventeen of the 100 elements were not flanked by identifiable TSDs, indicating that about 17% of the 1,380

elements might be false positives (see Methods and Supplementary Tables 1 and 2 for results and criteria used to identify TSDs). This value is probably an overestimate because the TSDs of old Pack-MULEs might have sustained mutations making them unrecognizable, or TSDs might have been deleted owing to 'abortive transposition'^{15,16}. Taken together, the results indicate that the vast majority of the original search output is Pack-MULEs.

Structures of representative Pack-MULEs are shown in Fig. 1, in which all elements shown have similar TIRs (all belong to the same RECON family) but sequences between the TIRs are largely different and include putative genes (sequences highly similar to known genes; Fig. 1a, b) and 'unknown' proteins (sequences matching cDNAs but not known genes; Fig. 1c). Collectively, the protein hits of the 1,380 Pack-MULEs involve diverse cellular functions including metabolism, transcription, cell defence and signal transduction, indicating that a wide variety of sequences have been captured.

Given the stringent criteria used, it is likely that genuine Pack-MULEs were missed and that there are significantly more than 1,380 Pack-MULEs in the genome. To estimate how many elements were missed by the genome-wide analysis, a second search was conducted of chromosomes 1 and 10, for which the sequences are nearly complete and systematic annotation is available^{17,18}. In this search, TIRs separated by more than 5 kb and MULEs with other repeats inserted in their internal sequence were not excluded. Instead, insertions were eliminated manually, and the remaining MULE sequences were subjected to the protein database search as described above, followed by examination of flanking sequences for the presence of a TSD (see Supplementary Tables 3 and 4 for TSDs). Using these criteria, 475 Pack-MULEs were found on the 65.7 Mb of chromosomes 1 and 10 (Fig. 2; Supplementary Tables 3 and 4). Of these, only 172 were identified in the genome-wide search, indicating that more than 60% $((475 - 172)/475 = 64\%)$ were missed.

Correcting for false positives (17%) and missed elements (64%) results in an estimate of 3,200 Pack-MULEs $(1,380 \times (100 - 17) \div (100 - 64))$ in 440 Mb of genome sequence. This assumes that the distribution of Pack-MULEs on chromosomes 1 and 10 is representative of those in the rest of genome. The distributions of the 319 Pack-MULEs on chromosome 1 and the 156 Pack-MULEs on chromosome 10 are shown in Fig. 2. Like the

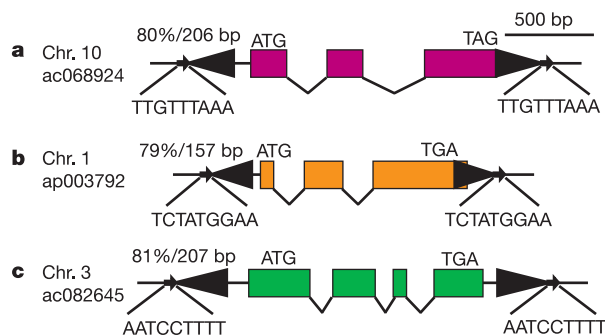


Figure 1 Pack-MULEs in Nipponbare. **a**, Putative peptide transporter; **b**, mitogen activated protein kinase-like protein; **c**, unknown protein. Pack-MULE TIRs are shown as black arrowheads, and black horizontal arrows indicate TSDs, with their sequences shown underneath. Exons are depicted as coloured boxes and introns as the lines connecting exons; other sequences are shown as horizontal lines. The GenBank accession number of the PAC or BAC sequence where the MULE is found and the chromosome where it is located are indicated. All elements depicted here have TIRs with high sequence identity. The gene prediction was furnished by the corresponding sequence provider. The similarity (percentage) and the length of the TIR of each element are also shown.

rice genes annotated on these chromosomes^{17,18}, Pack-MULEs are less abundant in the heterochromatic regions around centromeres and on the short arm of chromosome 10. Despite this uneven distribution, the average density of Pack-MULEs on these chromosomes is comparable (7.4 per Mb for chromosome 1 versus 7.0 per Mb for chromosome 10; see Supplementary Tables 3 and 4). If on average there are 7.2 Pack-MULEs per Mb, there would be about 3,100 Pack-MULEs in the genome ($7.2 \times 430 = 3,100$).

Among the 475 Pack-MULEs identified on chromosomes 1 and 10, 94 (20%) have two or more copies on the two chromosomes. These paralogous Pack-MULEs were probably generated by transposition rather than by large-scale genome duplication because in all cases distinct TSDs flanked duplicate elements (Supplementary Table 5). The genomic copy number of individual Pack-MULEs was estimated by randomly selecting 100 of the 475 Pack-MULEs on chromosomes 1 and 10 and using them to search rice genomic sequence for the presence of elements with both similar TIRs and internal regions (see Methods). Of the 100 Pack-MULEs, 27 are single-copy elements, 63 have two to five copies, and 10 have more than five copies (ranging from 6 to 14). With an average of three copies in the genome, most Pack-MULEs are low-copy-number elements.

To investigate the origin of the sequences captured by Pack-MULEs, the 100 randomly chosen Pack-MULEs from chromosomes 1 and 10 were used to query the rice genome for similar sequences that are not flanked by MULE TIRs. Among the 100 Pack-MULEs, 80 have one or more significant homologues (BLASTN $E < 10^{-10}$) (Supplementary Table 6) with a DNA sequence identity of 78–100% (mean 91%). Assuming that the captured sequence and the genomic homologue were initially identical, sequence acquisition must have occurred over a long time frame, perhaps millions of years, and might still be occurring. Of the 80 Pack-MULEs with identifiable genomic homologues, 73 have one or more homologues corresponding to the coding region of a predicted gene, and 46 have captured genomic sequences that are represented in cDNA collections (Supplementary Table 6). Most of these captured sequences are probably gene fragments, not complete genes, because the average length of captured fragments is only 325 bp (from 47 to 986 bp; Fig. 3, Supplementary Table 6). In addition, sequence acquisition by Pack-MULEs seems to occur at the DNA level as indicated by the conservation of introns between Pack-MULEs and their genomic homologues (where the corresponding cDNA sequences are available; see below and Fig. 3a, c).

There are several possible fates for the sequences in the remaining

Pack-MULEs (20 of 100) that display significant similarity to proteins in the database but not to rice genomic DNA. Acquired sequences might be from unsequenced regions (for example, heterochromatic regions), from rapidly evolving sequences or from sequences lost from the genome after capture.

From an evolutionary standpoint, of potentially greater interest is the discovery that Pack-MULEs can acquire genomic fragments from two or more chromosomal loci (Fig. 3). Of the 80 (out of 100) Pack-MULEs with identified genomic homologues, 18 (23%) contain sequences from two or more loci. The significant number of elements with sequences acquired from multiple loci led to a concern that such elements might be an artefact of sequence assembly errors. However, assembly errors would not generate elements that are flanked by a TSD. In this search all of the 18 chimaeric elements have a TSD. In addition, the chimaeric structure of several Pack-MULEs could be verified by the identification of the corresponding cDNA (see below, Fig. 3a, b and Supplementary Fig. 1D, E, G, H). Assembly error was also ruled out by PCR amplification of fragments of the predicted size from Nipponbare genomic DNA by using primers designed to amplify two overlapping fragments of each of 13 Pack-MULEs and flanking sequences (Supplementary Methods, Supplementary Table 7 and Supplementary Fig. 2).

Sequence divergence between captured gene fragments and genomic homologues permitted a determination of whether captured fragments are expressed (Supplementary Table 8). A search of the 475 elements on chromosomes 1 and 10 revealed that 25 (5%) are transcribed, on the basis of exact matches with rice full-length cDNAs¹⁹. Among the 34 cDNAs that correspond to the 25 Pack-MULEs (seven elements having two or more cDNA matches; Supplementary Table 8), the 5' ends of 23 (68%) are located within or adjacent to (within 40 bp) the TIR, and 8 (23%) were mapped to other parts of the element. Only three (9%) transcripts seem to be due to transcriptional readthrough, because their 5' ends were mapped to sequences flanking the Pack-MULE. Thus, the vast majority of Pack-MULE transcription is initiated from promoters in element sequences; either by promoters in the TIR or in acquired genomic sequence.

As another measure of the expression of captured DNA, the 475 Pack-MULEs from chromosomes 1 and 10, and the 1,380 Pack-MULEs from the genome-wide search, were used to query a database of 2,528 unique rice peptides²⁰. Although this database might represent less than 5% of rice proteins, six perfect matches were detected, with no other identical matches in the genome (see Methods and Supplementary Table 9).

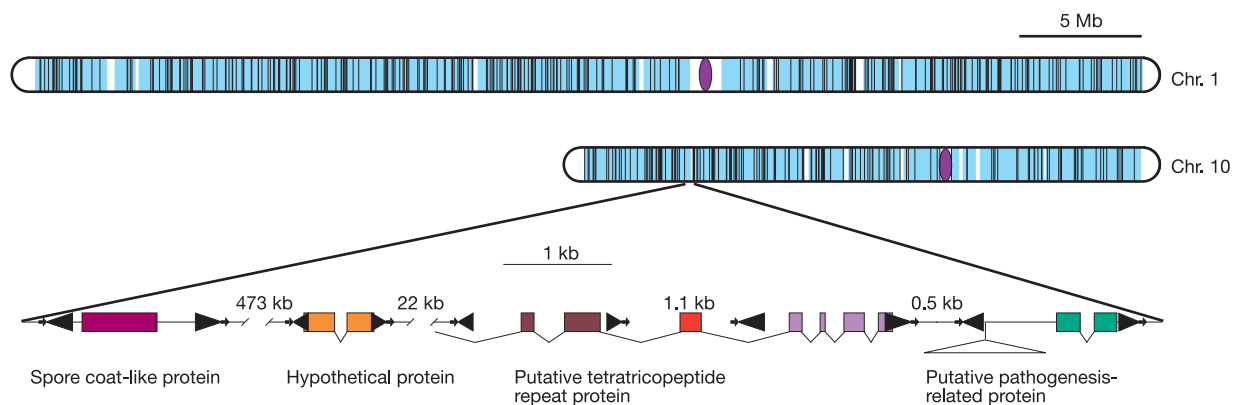


Figure 2 Distribution of Pack-MULEs on chromosomes 1 and 10. Blue blocks represent sequenced regions, whereas white spaces indicate gaps. Individual Pack-MULEs are represented by a vertical black line, and the purple ovals locate the centromeres. Below

chromosome 10 a region has been expanded, and individual Pack-MULEs are depicted as described in Fig. 1. Distances between adjacent Pack-MULEs are noted. A TE insertion in a Pack-MULE is noted with an open triangle.

Recognizing that rice expression libraries are incomplete, we sought an independent measure of gene function. To this end, ratios of the non-synonymous (K_a) to synonymous (K_s) substitution rates were calculated for the internal sequences of the randomly selected Pack-MULEs on chromosomes 1 and 10 (see above) and their genomic homologues. This analysis was restricted to the 54 sequence pairs (MULE versus genomic homologue; DNA sequence similarity 80–98%) sharing more than 50 amino acids. A K_a/K_s ratio significantly less than 1 suggests that the function of the acquired sequence either is being maintained or has been maintained for a significant period since acquisition. Of the 54 pairs, 18 have a ratio significantly smaller than 1 ($P < 0.05$; see Methods,

Supplementary Table 10). Theoretically, a low K_a/K_s might be due to the erroneous alignment of an element with a gene that is a paralogue of the gene acquired by the Pack-MULE. However, the results of a control experiment designed to test for such artefacts indicated that this was unlikely (see Supplementary Methods for details) and that the low K_a/K_s might reflect a functional constraint. Because the 54 sequence pairs were chosen from 100 Pack-MULEs and 5% of them might have a low K_a/K_s simply by chance ($P < 0.05$), the data suggest that more than 10% (18% – 5%) of the Pack-MULE sequences might have been functionally constrained despite the observation that the vast majority of captured DNA are gene fragments that are often rearranged (Fig. 3B, Supplementary Fig. 1B, 1C, 1E, 1H).

In summary, we report that there are over 3,000 Pack-MULEs in the rice genome with an average copy number of three. Because some elements contain two or more fragments, we estimate that more than 1,000 different gene fragments have been captured by MULEs. Although the mechanism of sequence capture is not yet understood, it is likely to involve the acquisition of genomic DNA rather than the cDNA copies of cellular transcripts as reported previously for human L1 retrotransposons^{21,22}. Furthermore, about one-fifth of the identified Pack-MULEs contain fragments acquired from multiple genomic loci, thus demonstrating their potential to create novel genes through the duplication, rearrangement and fusion of diverse genomic sequences. □

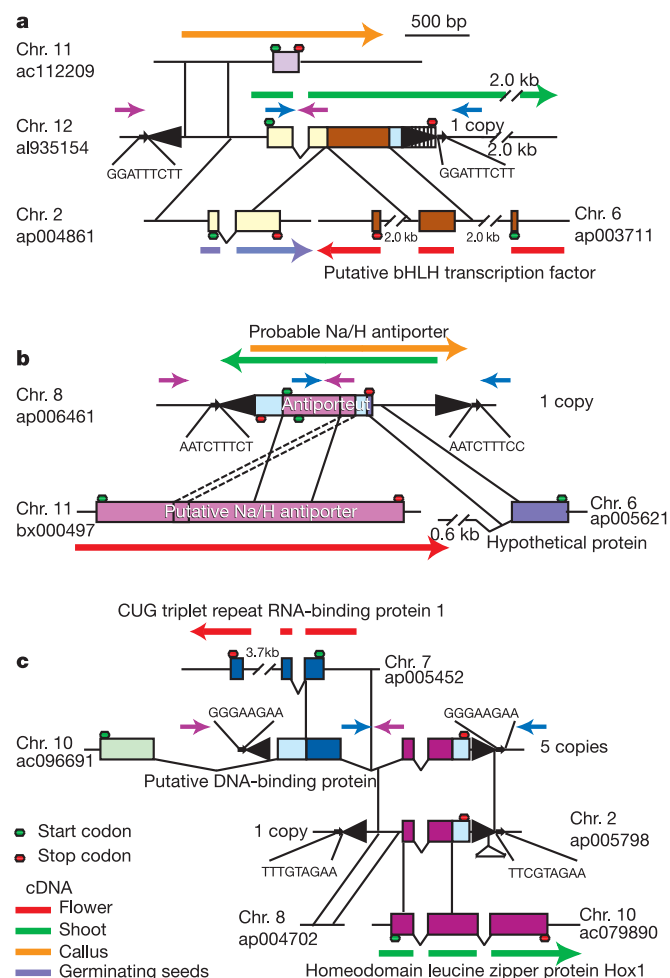


Figure 3 Structure and genomic origin of chimaeric Pack-MULEs. **a**, A Pack-MULE containing gene fragments from three genomic loci including one intron. **b**, A Pack-MULE with a new ORF derived from a rearranged Na/H antiporter gene. **c**, Possible step-wise formation of a chimaeric Pack-MULE: a Pack-MULE on chromosome 10 with sequences acquired from three loci (on chromosomes 7, 8 and 10) and an apparent intermediate element (on chromosome 2) with the gene fragments from chromosomes 8 and 10. Homologous regions are associated with solid or dashed lines. Light-blue boxes represent exons (or part of an exon) where the origin of the sequence is not clear. The striped box in **a** indicates that the TIR overlaps with the putative exon. Long coloured arrows indicate sequences matching cDNAs from the designated tissues. Small arrows (purple and blue) over the Pack-MULEs indicate the location of the primers used in PCR amplification of Pack-MULE fragments (see the text, Supplementary Methods, Supplementary Table 7 and Supplementary Fig. 2 for details). The gene name is given for putative genes and hypothetical proteins; all other genes encode 'unknown proteins'.

Methods

Computer-assisted identification of repeats in rice

Sequences of repeat families in rice were identified with RECON (version 1.03) as described^{11,23}, with 400 Mb of Nipponbare genomic sequence (downloaded from <http://rgp.dna.affrc.go.jp> on 23 August 2002) for the initial all-versus-all comparison. The resulting 3,300 repeat families (within each family, more than 90% of the sequence can be aligned between any two members on the basis of the shorter sequence) were examined individually and those derived from TEs were analysed further. MULEs were identified by features of known MULEs, including the presence of TIRs and TSDs (see Supplementary Methods for details). Curated TE sequences, including MULE TIRs, were used to mask the 440 Mb of Nipponbare sequence (downloaded from <http://rgp.dna.affrc.go.jp> on 7 January 2003) with RepeatMasker (version 07/07/2001, using default parameters; <http://ftp.genome.washington.edu/RM/webrepeatmaskerhelp.html>). RepeatMasker output files contain annotations of all sequences that matched TEs as well as their position in the input genomic sequence and are the basis for both the genome-wide search and the search of chromosomes 1 and 10. Specifically, a putative MULE must have both ends hit by the same RECON TIR family using RepeatMasker (see Supplementary Methods and Supplementary Table 1 for how a MULE TIR is defined from RECON output). The size of Pack-MULE TIR consensus ranges from 80 to 496 bp, and the similarity of MULE TIRs range from 68% to 98%). Before the search, the RepeatMasker output was further processed so that only the best hit remained when the same locus was hit by several TE consensus sequences. The relative position of MULE and other TEs (for example, whether there are additional TEs between TIRs) was also determined using the output.

The genome-wide search of Pack-MULEs

In the genome-wide search, all possible pairs of annotated MULE TIRs located within 5 kb of each other (with no additional TEs between the TIR) were examined by following the criteria listed in the text. MULEs recovered in this way included both Pack-MULEs and other MULEs that do not contain cellular genes (or gene fragments). To distinguish the two, BLASTX was used (with the option `wordmask = seg`) to identify MULEs containing sequences that are similar (with $E < 10^{-9}$) to proteins in GenBank or to the annotated proteins of the 93-11 (*indica*) genome (downloaded from [ftp://ftp.genome.org/pub/grame/protein/sequence/Os_indica_protein_BGL.txt](http://ftp.genome.org/pub/grame/protein/sequence/Os_indica_protein_BGL.txt) on 17 April 2003). For each MULE, the best protein hit was excluded because it might have been a self-hit. Those MULEs with hits to non-hypothetical, non-transposase proteins were defined as Pack-MULEs.

TSDs and copy number estimates for individual Pack-MULEs

The length of Pack-MULE TSDs ranges from 7 to 11 bp. The most common TSD is 9 bp (417 out of 475, or 88%; Supplementary Tables 2, 3 and 4). For TSDs that are 9 bp or longer, a maximum of two mismatches (or one mismatch plus one single-base indel) was allowed. For TSDs that are 8 bp and 7 bp, a maximum of one mismatch or one single-base indel was allowed (see Supplementary Tables 3 and 4 for all TSD sequences). For individual Pack-MULEs, if the TIRs of the two elements belonged to the same family identified by RECON, and more than 50% of the sequence between the TIRs could be aligned (BLASTN, $E < 10^{-10}$), the two elements were defined as copies that arose from the same element.

Gene prediction

If there was a cDNA match, gene predictions (except in Fig. 3a; see below) were based on the annotation provided by the rice full-length cDNA consortium¹⁹. If not, the annotation used was from the corresponding sequence provider or the annotation of rice contigs (version 07232003) from the Institute of Genomic Research (TIGR) (<http://www.tigr.org>). In Fig. 3a, the open reading frame (ORF) inside the Pack-MULE was defined by ORFfinder (<http://www.ncbi.nlm.nih.gov>) on the basis of the corresponding cDNA sequence¹⁹.

Expression analysis

The full-length cDNA data set was downloaded from <http://cdna01.dna.affrc.go.jp/cDNA/> on 15 November 2003. A Pack-MULE is considered to have a cDNA match if, first, sequence similarity between the Pack-MULE and the cDNA is higher than 99.5% over the entire length of the cDNA and, second, the chromosomal position of the Pack-MULE is consistent with the genomic position of the particular cDNA provided by the rice full-length cDNA consortium¹⁹. Peptide sequences representing 2,528 unique proteins were downloaded from ref. 20 (<http://www.pnas.org>) and used as queries to search against all Pack-MULE sequences recovered in both the genome-wide search and the search of chromosomes 1 and 10 with TBLASTN²⁰. Peptide sequences that generated perfect hits with Pack-MULE sequences were then used to search the rice genomic database. A particular Pack-MULE sequence was considered to have a peptide match if it was the only perfect hit in the genome.

K_a/K_s analysis

The sequences of each Pack-MULE and its corresponding genomic homologue were aligned using the 'pileup' program from the University of Wisconsin GCG program suite (version 10.1) accessed through Research Computing Resources at the University of Georgia. The alignment was based on the ORF of the genomic homologue (see Supplementary Tables 6 and 10 for details). If the genomic homologue was not predicted as a coding sequence, the ORF of the MULE was used. K_a and K_s were calculated with MEGA using the Pamilo-Bianchi-Li method²⁴. The significance of purifying selection (P value) was evaluated with the z -test in MEGA.

Received 9 June; accepted 13 August 2004; doi:10.1038/nature02953.

1. Lisch, D. *Mutator* transposons. *Trends Plant Sci.* **7**, 498–504 (2002).
2. Walbot, V. & Rudenko, G. N. In *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 533–564 (American Society of Microbiology Press, Washington, D.C., 2002).
3. Chalvet, F., Grimaldi, C., Kaper, F., Langin, T. & Daboussi, M. J. Hop, an active *Mutator*-like element in the genome of the fungus *Fusarium oxysporum*. *Mol. Biol. Evol.* **20**, 1362–1375 (2003).
4. Talbert, L. E. & Chandler, V. L. Characterization of a highly conserved sequence related to *mutator* transposable elements in maize. *Mol. Biol. Evol.* **5**, 519–529 (1988).
5. Yu, Z., Wright, S. I. & Bureau, T. E. *Mutator*-like elements in *Arabidopsis thaliana*. Structure, diversity and evolution. *Genetics* **156**, 2019–2031 (2000).
6. Turcotte, K., Srinivasan, S. & Bureau, T. Survey of transposable elements from rice genomic sequences. *Plant J.* **25**, 169–179 (2001).
7. Benneken, J. L. & Springer, P. S. The generation of *mutator* transposable element subfamilies in maize. *Theor. Appl. Genet.* **87**, 657–667 (1994).
8. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).
9. Normile, D. & Pennisi, E. The rice genome. Rice: boiled down to bare essentials. *Science* **296**, 32–36 (2002).
10. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
11. Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
12. Cresse, A. D., Hulbert, S. H., Brown, W. E., Lucas, J. R. & Benneken, J. L. *Mu1*-related transposable elements of maize preferentially insert into low copy number DNA. *Genetics* **140**, 315–324 (1995).
13. Hanley, S. *et al.* Identification of transposon-tagged genes by the random sequencing of *Mutator*-tagged DNA fragments from *Zea mays*. *Plant J.* **23**, 557–566 (2000).
14. Le, Q. H., Wright, S., Yu, Z. & Bureau, T. Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **97**, 7376–7381 (2000).
15. Taylor, L. P. & Walbot, V. A deletion adjacent to the maize transposable element *Mu-1* accompanies loss of *Adh 1* expression. *EMBO J.* **4**, 369–376 (1985).
16. Levy, A. & Walbot, V. Molecular analysis of the loss of somatic instability in the *bz2-mu1* allele of maize. *Mol. Gen. Genet.* **229**, 147–151 (1991).
17. Sasaki, T. *et al.* The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312–316 (2002).
18. The Rice Chromosome 10 Sequencing Consortium. In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* **300**, 1566–1599 (2003).
19. Kikuchi, S. *et al.* Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* **301**, 376–379 (2003).
20. Koller, A. *et al.* Proteomic survey of metabolic pathways in rice. *Proc. Natl Acad. Sci. USA* **99**, 11969–11974 (2002).
21. Moran, J. V., DeBerardinis, R. J. & Kazanian, H. H. Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
22. Emerson, J. J., Kaessmann, H., Betran, E. & Long, M. Extensive gene traffic on the mammalian X chromosome. *Science* **303**, 537–540 (2004).
23. Jiang, N. *et al.* An active DNA transposon in rice. *Nature* **421**, 163–167 (2003).
24. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. *MEGA2: Molecular Evolutionary Genetics Analysis software* (Arizona State Univ., Tempe, Arizona, 2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank I. K. Jordan and E. K. Kentner for assistance with K_a/K_s analysis, R. Liu for verification of TSD sequences, C. Feschotte for advice, and J. Benneken and E. Pritham for critical reading of the manuscript. This study was supported by a grant from the NSF Plant Genome Program to S.R.W. and S.R.E. N.J., X.Z. and S.R.W. were also supported in part by grants from the NIH, and Z.B. and S.R.E. were also supported in part by grants from the NIH, the Howard Hughes Medical Institute, and Alvin Goldfarb.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.R.W. (suew@plantbio.uga.edu).

Small modulation of ongoing cortical dynamics by sensory input during natural vision

József Fiser^{1,3}, Chiayu Chiu² & Michael Weliky^{1,3}

¹Center for Visual Science, ²Interdepartmental Program in Neuroscience, and ³Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York 14627, USA

During vision, it is believed that neural activity in the primary visual cortex is predominantly driven by sensory input from the environment. However, visual cortical neurons respond to repeated presentations of the same stimulus with a high degree of variability^{1–4}. Although this variability has been considered to be noise owing to random spontaneous activity within the cortex^{5–7}, recent studies show that spontaneous activity has a highly coherent spatio-temporal structure^{8–13}. This raises the possibility that the pattern of this spontaneous activity may shape neural responses during natural viewing conditions to a larger extent than previously thought. Here, we examine the relationship between spontaneous activity and the response of primary visual cortical neurons to dynamic natural-scene and random-noise film images in awake, freely viewing ferrets from the time of eye opening to maturity. The correspondence between evoked neural activity and the structure of the input signal was weak in young animals, but systematically improved with age. This improvement was linked to a shift in the dynamics of spontaneous activity. At all ages including the mature animal, correlations in spontaneous neural firing were only slightly modified by visual stimulation, irrespective of the sensory input. These results suggest that in both the developing and mature visual cortex, sensory evoked neural activity represents the modulation and triggering of ongoing circuit dynamics by input signals, rather than directly reflecting the structure of the input signal itself.

Using an implanted multi-electrode array, we measured population activity within the visual cortex of awake, freely viewing ferrets at three different developmental stages: immediately after eye opening at postnatal day (P) 30–32 ($n = 3$), immediately after the maturation of orientation tuning and long-range horizontal connections at P44–45 ($n = 3$), and in the mature animal at P83–90 ($n = 4$). A linear array of 16 microwire electrodes, spanning 9.0 mm, was placed at 300–500- μ m depth in layer 2/3 of the striate cortex, which allowed recordings to be obtained from cortical sites with varying degrees of receptive field overlap (Fig. 1a). Fifteen 100-s recording trials were acquired in head-restrained ferrets under each of three interleaved conditions: (1) presentation of dynamic natural scenes from a film; (2) presentation of dynamic random-noise stimuli; and (3) complete darkness (see Methods). At all ages,

both natural-scene and random-noise visual stimulation increased the mean and variance of neuronal firing rates as well as the percentage of active bins (that is, bins containing at least one spike) when compared with the same measures in the dark (Fig. 1b). Thus, both types of stimuli were able to drive neurons within the visual cortex.

Using a statistical approach in which the input signal and output neural responses were characterized by measuring their ensemble properties, we assessed the degree to which the activity of visual cortical neurons accurately reflects the spatio-temporal structure of the visual input. This was accomplished by computing temporal and spatial correlation functions for the dynamic natural-scene and random-noise stimuli, and comparing these with the same measures computed from neural activity recorded under the three experimental viewing conditions. Whereas correlation functions for the neural responses were computed from recorded multi-unit activity, those calculated for the dynamic natural-scene and random-noise stimuli were based on the response of a bank of linear filters applied to the two types of stimuli (see Methods). Such linear filters characterize the response properties of primary visual cortical neurons, which are strongly activated by changing high-contrast features, such as moving edges and junctions, and very weakly activated by low-contrast homogeneous image regions. Thus, the temporal and spatial correlation functions computed from these

filter responses provide a statistical measure of the expected activation of neurons at different recording sites by image features in the dynamic stimulus displays.

As expected, the correlation functions computed from filter responses to dynamic random-noise and natural-scene stimuli were very different, demonstrating the diverse statistical properties of the two displays (Fig. 1c). The temporal correlation function for natural scenes revealed strong correlations extending to several hundred milliseconds, reflecting the continuity of images across successive film frames. On the other hand, filter responses to random noise were decorrelated at 30 Hz, consistent with random changes across consecutive frames in the stimuli. The spatial correlation function for random noise revealed a gradual decrease in correlated activity as a function of increasing distance between recording sites, reaching zero between sites whose receptive fields did not overlap. For natural scenes, high correlations were present between all recording sites irrespective of the distance between sites, due to the extended contours of objects that span non-overlapping receptive fields.

If neural activity evoked by visual stimulation was directly related to the input signal, then the different statistical properties of natural-scene and random-noise displays would be expected to trigger different patterns of neuronal firing within the primary visual cortex. To quantify the degree of similarity between

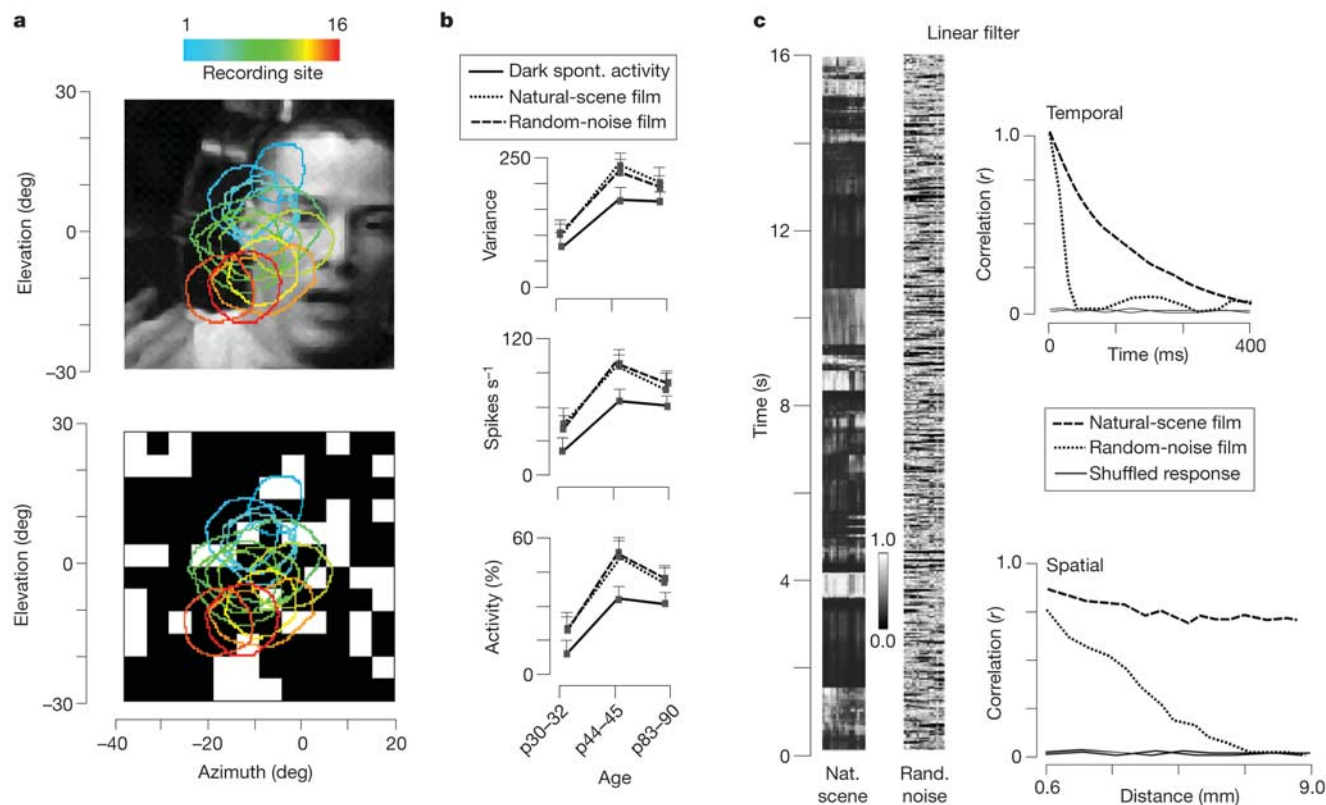


Figure 1 Statistical properties of natural-scene and random-noise film images.

a, Representative images from natural-scene film (top) and random-noise stimuli (bottom). The circles overlaid upon each image show mapped classical receptive fields for each of the 16 recording sites in a P83 ferret. Colour is used to distinguish the relative locations of receptive fields recorded at the different recording sites. The receptive fields for recording sites 1–3 and 14–16 are completely non-overlapping. A similar distribution of overlapping receptive fields was obtained in all younger P30–32 and P44–45 visual cortices. **b**, The variance of neuronal firing in spikes s^{-2} (top), the firing rate of neurons (middle), and the percentage of active spike bins (bottom) under the different stimulus conditions at all developmental ages. Bin width, 25 ms. Error bars represent s.e.m. **c**, Response of 16 linear filters to natural-scene and random-noise film images.

Orientation and spatial frequency tuning, as well as receptive fields and the impulse response functions of linear filters were derived from recordings obtained in a P45 ferret. Similar results were obtained from ferrets at all other developmental ages. Left: time series plots of linear filter responses to stimulus films. Columns represent the simulated activity of each of the 16 recording sites. In response to the natural-scene film, periods of strong activity (white), reflecting movement of high-contrast edges through the receptive fields, are separated by periods of weak activity (black), when no stimulus features drive the filters. Random-noise film images stimulate the filters more uniformly. Right: temporal and spatial correlation functions of filter responses for natural-scene and random-noise film images. Shuffled response refers to the control condition of randomized filter responses.

correlation functions computed for neural activity during the different viewing conditions, a linear regression coefficient was fitted to an initial segment of each correlation function (temporal correlation function, 50–200 ms; spatial correlation function, 0.6–3.0 mm). At P30–32, visual cortical neurons burst in a slow and irregular manner in the absence of and during visual stimulation (Fig. 2). As a result, identical temporal correlations extending to several hundred milliseconds and identical patterns of locally correlated activity between neighbouring cortical sites were observed under all three viewing conditions ($P > 0.05$, ANOVA) (Fig. 3a). During subsequent development, temporal correlations shortened while spatial correlations broadened for all viewing conditions. At P44–45, spatial correlation functions of neural activity during all three viewing conditions were still identical ($P > 0.05$, *t*-test). However, small differences in temporal correlations emerged between P30–32 and P44–45. Whereas temporal correlations in dark spontaneous activity and random-noise responses remained indistinguishable at P44–45 ($P > 0.05$, *t*-test), temporal correlations in natural-scene responses fell off at a slightly slower rate ($P < 0.05$, *t*-test).

With continued maturation of the visual cortex, spontaneous and visually evoked activity evolved to dominant synchronized oscillations by P83–90 (Figs 2 and 3b). Between P44–45 and P83–90, dark spontaneous activity and random-noise responses became increasingly temporally decorrelated (Fig. 3a), and temporal correlation functions under these two conditions remained indistinguishable ($P > 0.05$, *t*-test). Temporal correlations in natural-scene responses also became increasingly decorrelated between P44–45

and P83–90, but continued to be significantly different from the other two conditions ($P < 0.05$, *t*-test). As a result of these developmental shifts, temporal correlations in neural responses at P83–90 reflected the different statistical properties of the two stimuli: temporal correlations in natural-scene responses fell off at a slower rate than those of random-noise responses, which were rapidly decorrelated. The shift in neuronal firing patterns also led to changes in the spatial correlation functions. At P83–90, short- and long-range spatial correlations under all viewing conditions reached high levels that were nearly uniform between cortical sites regardless of their separation. The slopes of spatial correlation functions under dark and natural-scene viewing conditions were not significantly different ($P > 0.05$, *t*-test), and correlated activity was only 20% higher during viewing of natural scenes than in the dark. Notably, there was a significant difference between spatially correlated neural activity evoked by the two types of visual stimuli. For neighbouring cortical sites with overlapping receptive fields, correlations in neural activity during random-noise and natural-scene stimulation were equally high ($P > 0.05$, *t*-test). However, as the separation between cortical sites increased, the level of correlated activity fell off more steeply during random-noise viewing than during natural-scene viewing ($P < 0.05$, *t*-test). This difference in slope reflected the different spatial properties of the two stimuli: long-range correlations in natural scenes are significantly higher than those in random noise. These results demonstrate that, as a result of developmental shifts in spatial and temporal correlation functions, there was a significantly improved correspondence between visually evoked neural activity and the structure of the input signal with increasing age.

It is important to emphasize that even in the mature visual cortex, visually evoked changes in the rate and pattern of neural firing were much smaller than expected if neurons were primarily driven by sensory input. Although visual stimulation increased neural activity in the cortex, the firing rates of neurons during visual presentation increased by only 23% compared with those in the dark. At 50 ms, the difference in temporal correlation coefficients of natural-scene and random-noise displays was 0.47, as measured by the linear filter model, whereas the actual difference in evoked neural activity to the two stimuli was only 0.039 ± 0.008 . Furthermore, spatial correlations evoked by random-noise stimuli did not fall off to zero at large cortical separations as predicted by the linear filter responses. Rather, the correlation coefficient dropped by only 20% to the mean level of spatial correlations present in ongoing activity. These modest changes were not merely due to averaging, as demonstrated by individual assessment of mean correlations in neuronal firing between each pair of recording sites (see Supplementary Information). At all ages, such pairwise analysis generated correlation patterns that were nearly identical under the different viewing conditions ($P > 0.05$, ANOVA). Taken together, these findings suggest that during natural viewing, spatio-temporal correlations in neural firing are not primarily governed by the statistical properties of the input signal, but rather are dominated by the correlational structure of ongoing activity.

Eye movements, or lack of visual stimulation, could be a potential cause of the reduced correspondence between evoked neural responses and the input statistics. To test this possibility, we monitored eye movements during random-noise and natural-scene presentation. Temporal and spatial correlations, re-computed after excluding portions of binned spike traces recorded when the eyes were closed or during saccadic eye-movements, were not significantly different from correlation functions computed using the full spike traces ($P > 0.05$, *t*-test) (see Supplementary Information). Thus, eye movement artefacts or inadequate visual stimulation cannot explain the relatively small modulation of spontaneous activity by the visual input.

The prolonged neuronal bursting observed in younger animals, which was responsible for the strong temporal correlations at these

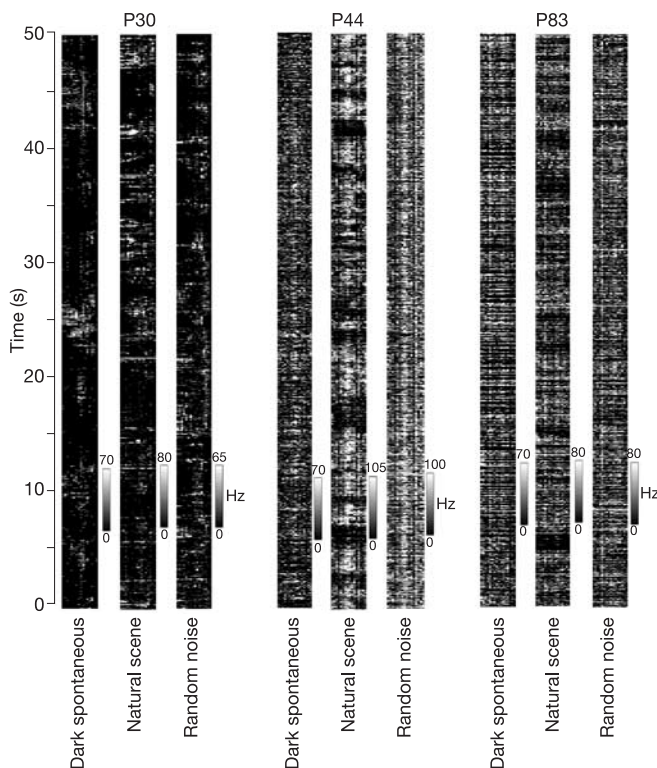


Figure 2 Time series plots of neural activity recorded under the three interleaved stimulus conditions at three different ages. The time series graphs were obtained from a single animal in each age group. At each age, visual stimulation modulates the spatio-temporal pattern of spontaneous activity, but does not significantly alter its basic correlational structure. At P44 and P83, the continuous spontaneous activity present across all recording sites is modulated during specific periods by the natural-scene film, presumably when a high-contrast feature moves across the receptive fields. The random-noise film image does not induce such significant modulations in firing rate. Bin width, 20 ms.

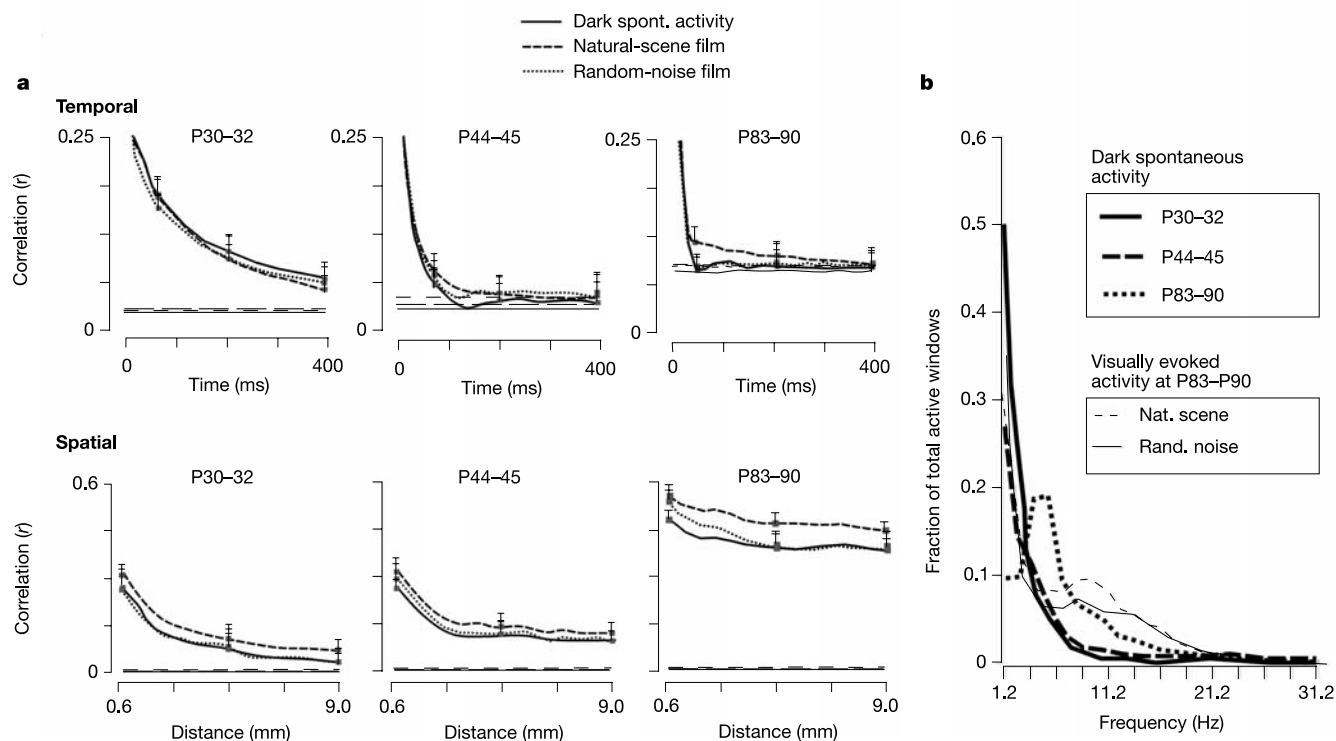


Figure 3 Developmental changes in the spatio-temporal pattern of stimulus-evoked and spontaneous visual cortical activity in awake-behaving ferrets. **a**, Correlation functions computed for dark spontaneous activity, as well as evoked activity to natural-scene and random-noise films in awake ferrets at three different ages. Thin horizontal lines show plots of correlation functions computed at each age for each condition using randomly shuffled binned spikes. Random temporal shuffling of spike trains abolished all correlations in all three age groups, demonstrating that the observed shifts in correlated activity were not simply a result of the developmental increase in cell firing rates. Top row:

temporal correlation functions. Bin width, 20 ms. Bottom row: spatial correlation functions. Bin width, 50 ms. Error bars represent s.e.m. **b**, Emergence of oscillations in dark spontaneous activity during visual cortical development. Dominant 4–8-Hz oscillations emerge between P44–45 and P83–90. At P83–90, visual stimulation abolishes oscillations at these frequencies but triggers higher frequency oscillations. Total active windows are the total number of 800-ms windows that contained at least 1 spike.

ages, could be caused by either an intrinsic inability of neurons in the immature visual cortex to respond rapidly to visual stimuli, or by reverberating neural activity within immature feedback circuits¹⁴. To distinguish between these two possibilities, recordings were obtained from each animal under light anaesthesia using the same visual stimulation conditions as in the awake-behaving context (Fig. 4a). Light anaesthesia significantly dampened neuronal excitability, as demonstrated by a reduction in spontaneous activity, whereas neuronal responsiveness to visual stimulation was maintained (see Supplementary Information). In contrast to awake animals, temporal correlations, both during and in the absence of visual stimulation, showed swift decorrelation by 50 ms at all ages, including the youngest P30 animals (Fig. 4b, top). Thus, the slope of temporal correlation functions, computed for neural activity between 50–200 ms under all three stimulus conditions, was significantly reduced at P30–32 and P44–45 ($P < 0.05$, t -test) to levels matching those for control P83–90 animals (see Supplementary Information). This indicates that cortical neurons are capable of rapidly responding to sensory signals at all ages, and suggests that temporal correlations extending to several hundred milliseconds in the young awake animals may result from recurring activity generated within immature feedback circuits. In contrast, the spatial correlation functions were not substantially different for the awake and anaesthetized animal at any age (compare Fig. 4b, bottom, with Fig. 3a, bottom, $P > 0.05$, t -test). This suggests that feed-forward sensory evoked activity, which is maintained during anaesthesia, has a prominent role in governing the correlated firing of neurons between different cortical sites.

The strong correspondence between patterns of correlated

neuronal activity during and in the absence of visual stimulation, indicates a tight relationship between spontaneous activity and the cortical representation of sensory signals. Even when stimulated by input signals with diverse statistical properties, the firing patterns of visual cortical neurons are dominated by the intrinsic dynamical properties of the cortical circuit rather than the signal statistics. These results agree with previous reports in anaesthetized preparations and *in vitro* brain slices demonstrating that sensory evoked responses are significantly modulated by ongoing spontaneous neuronal activity, and that this spontaneous activity reflects the intrinsic dynamical behaviour of neural circuits^{13,15–17}. They also agree with previous studies in nonhuman primates showing that the firing rates of visual cortical neurons are modulated to a smaller degree during free viewing of natural scenes than when the same stimuli are statically flashed within the classical receptive field during fixation^{18,19}. However, our results are the first to demonstrate that intrinsic circuit dynamics strongly govern the correlated firing of neurons in primary visual cortex of freely viewing animals. These results provide the most direct evidence that the intrinsic dynamical behaviour of neural circuits might have a prominent role in visual sensory coding. If the specific spatio-temporal structure of these dynamics is important for normal sensory processing, then the shift from slow asynchronous bursting at early ages to dominant synchronous 4–8-Hz oscillations in the adult probably has a central role in the emergence of sensory coding during brain development.

Our analysis has focused only on second-order spatio-temporal correlations in spontaneous and visually driven activity, and it does not address higher-order correlations in neural activity that may also be used to encode input signals^{20–22}. In addition, our results are

based on recording in the upper layers of the primary visual cortex. It is conceivable that cell responses in layer 4 are more readily related to the structure of the input stimulus^{15,23,24}. Because of the high-energy consumption of baseline neural activity in the brain^{25,26}, it would be inefficient to maintain the observed high level of spon-

taneous activity unless it had an essential role in sensory processing. We propose that during sensory coding, stimulus-evoked activity in the visual cortex principally reflects the modulation and triggering of intrinsic circuit dynamical behaviour by sensory signals, instead of directly encoding the structure of the input signal itself. In this framework, ongoing activity may not be noise upon which visual responses are superimposed, but rather an integral component of sensory processing. □

Methods

Electrode implant procedure

Surgical procedures were identical to those described previously¹². Briefly, anaesthesia was induced and maintained during surgery by inhalation of isoflurane (0.5–2.0%) in a 2:1 nitrous oxide/oxygen mixture. A section of skull was exposed over area 17, the dura was reflected and the electrode array aligned along the caudal bank of the posterior lateral gyrus. After lowering the electrode array so that it touched the cortical surface, the exposed brain was covered by agar, and a headset was affixed by dental acrylic to the skull. A separate headpost holder was attached to the frontomedial skull by stainless steel screws and dental acrylic. All procedures were approved by the University of Rochester Committee on Animal Research.

Recording and data acquisition

The multi-electrode array consisted of a single row of 16 electrodes spaced at 600 μm (9.0-mm wide). Each electrode was a 12.5- μm -diameter tungsten wire with 2.5- μm H-M-L insulation (California Fine Wire). The insulation along the final 30–60- μm length of wire was removed, creating a 200–300 k Ω impedance electrode. The electrodes typically provided clear multi-unit signal on each channel with occasional single unit signal. The average noise amplitude was $5.1 \pm 0.9 \mu\text{V}$, whereas the average signal amplitude was $34.4 \pm 9.8 \mu\text{V}$. All electrodes were simultaneously raised or lowered by turning a single, 100-thread-per-inch screw, and they were connected to custom-made amplifiers providing a gain of 20,000. The signal was band-pass-filtered between 600 and 6,000 Hz and digitized at 10,000 kHz via an AD board (National Instruments) to a PC. Data acquisition was performed with custom-written Labview programs (National Instruments). Spike discrimination was done offline by manually setting a separate voltage threshold for each electrode. Stable recordings were maintained for 8–12 h.

Recordings were initiated after 2–3 h of recovery from anaesthesia, when the animal was fully alert. There were delays of approximately 10–20 s between interleaved trials. After awake recordings, the same sessions were repeated under light anaesthesia (0.5–1.0% isoflurane in a 2:1 nitrous oxide/oxygen mixture) and then the receptive field properties were measured. Heart rate was monitored and body temperature was maintained with a thermostatically controlled heating blanket. At each of 16 recording sites, the position and size of receptive fields were determined by a reverse correlation method using an array of flashed white squares 2–3° on each side, and orientation and spatial frequency tuning were assessed by flashed high-contrast sine-wave gratings²⁷. Stimuli were repeated at least eight times and evoked responses were averaged to obtain the receptive field structure and the temporal impulse response function. Drifting of receptive field positions across epochs was rarely encountered, and those epochs with receptive centres more than one square deviating from the mean were excluded from the analysis. Similar to earlier results, the receptive field sizes varied between 10–18° in size and were in the central 30° of the visual field at all three ages.

Visual stimulation

While the animal rested comfortably on a padded platform, the head was held in a fixed position with the use of a rigid metal post. The animal was free to make natural eye movements. A 4 × 3 foot stimulus screen, which was placed at a distance of 30 cm from the head, covered 130° by 100° visual angle. An LCD projector provided both random-noise and natural-scene stimuli, which covered the entire screen, at a resolution of 1,024 × 768 pixels and a refresh rate of 75 Hz. Random-noise stimuli were generated by a two-dimensional array of white squares, which were randomly flashed on a black background. For each frame, a square could appear with 25% uniform probability at each position in space, which occupied 5° × 5° visual angle. A new display was presented at a rate of 75 Hz. The trailer for the film *The Matrix* was used as the natural-scene film stimulus, which was presented at 24 Hz update rate and 720 × 480 resolution.

Correlation analysis

The temporal correlation function describes the mean correlation between spike firing rates measured at all recording sites at one particular time point, and successive time points. Discriminated spikes were placed into 20-ms bins. The Pearson product-moment correlation coefficient (r) was computed between population spike activity at each time bin t and successive time bins $t + n$. A histogram of the r -values at each $t + n$ interval was constructed across all time bins and all trials for each animal. The spatial correlation function describes the mean correlation r between the firing rates measured simultaneously at pairs of recording sites, parameterized by the distance between the recording sites.

Linear filter model

The spatial receptive fields of the linear filter model were constructed according to previously described procedures²⁷. Briefly, the spatial properties of the filters were defined by experimentally measured receptive field size, shape and position, as well as the

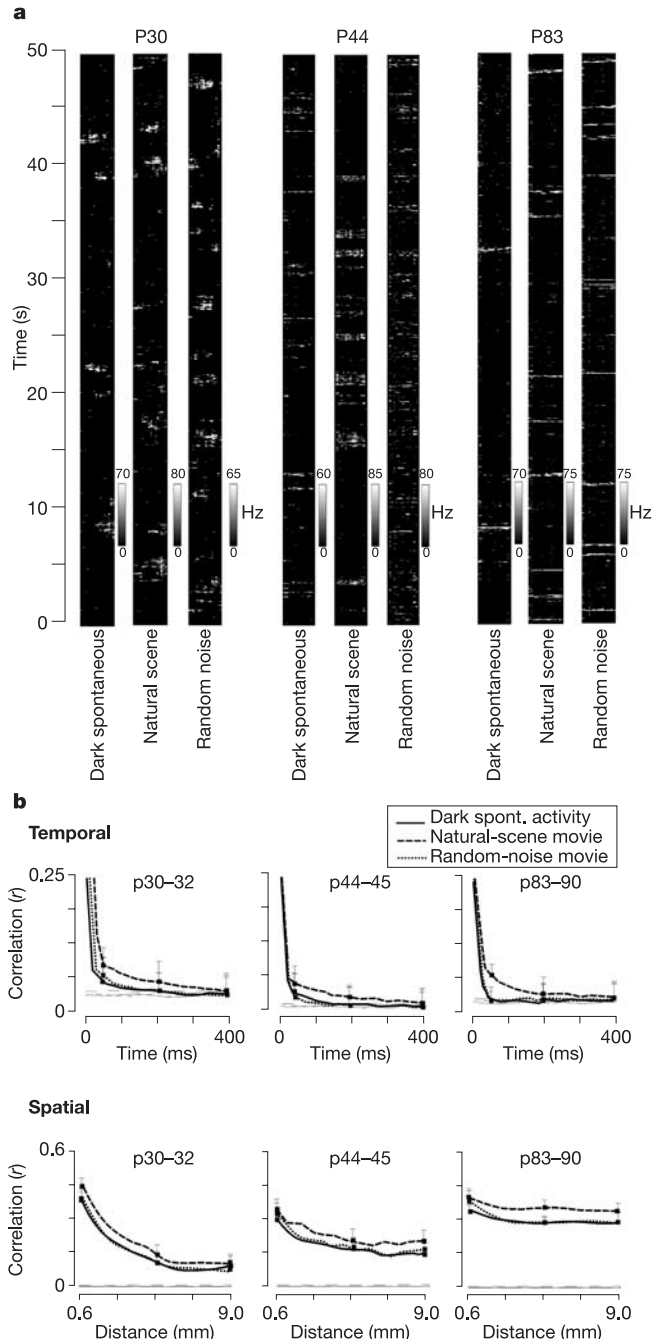


Figure 4 Developmental changes in the spatio-temporal pattern of stimulus-evoked and spontaneous visual cortical activity in anaesthetized ferrets. **a**, Comparison of time series plots for dark spontaneous activity, natural-scene and random-noise films at three different ages. Plots are from a single animal at each age under the same stimulus conditions as awake animals. Bin width, 20 ms. **b**, Correlation functions computed for dark spontaneous activity, as well as evoked activity to natural-scene and random-noise film images in anaesthetized ferrets. Thin horizontal lines show plots of correlation functions computed at each age, using randomly shuffled binned spikes. Top row: temporal correlation functions. Bin width, 20 ms. Bottom row: spatial correlation functions. Bin width, 50 ms. Error bars represent s.e.m.

orientation and spatial frequency tuning, at each recording site for each animal. The temporal impulse functions were derived by generating a histogram of the neural responses to briefly presented flashes at random positions within the receptive field. Identical results were obtained with filters defined by parameters that were based on data obtained at each of the three age groups. The shapes of the correlational functions were relatively insensitive to the particular parameter values of the spatial and temporal filters.

Oscillation analysis

Oscillation frequency was computed separately at each recording site. First, discriminated spikes were placed into 2-ms bins. 800-ms windows of binned spikes were extracted every 200 ms, and the autocorrelation of the windowed spikes was computed, followed by the fast Fourier transform (FFT). The frequency with the maximum amplitude was determined, and a histogram of these frequencies was constructed.

Received 3 May; accepted 30 July 2004; doi:10.1038/nature02907.

- Henry, G. H., Bishop, P. O., Tupper, R. M. & Dreher, B. Orientation specificity and response variability of cells in striate cortex. *Vision Res.* **13**, 1771–1779 (1973).
- Schiller, P. H., Finlay, B. L. & Volman, S. F. Short-term response variability of monkey striate neurons. *Brain Res.* **105**, 347–349 (1976).
- Vogels, R., Spileers, W. & Orban, G. A. The response variability of striate cortical neurons in the behaving monkey. *Exp. Brain Res.* **77**, 432–436 (1989).
- Azouz, R. & Gray, C. M. Cellular mechanisms contributing to response variability of cortical neurons in vivo. *J. Neurosci.* **19**, 2209–2223 (1999).
- Zohary, E., Shadlen, M. N. & Newsome, W. T. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143 (1994).
- Shadlen, M. N. & Newsome, W. T. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *J. Neurosci.* **18**, 3870–3896 (1998).
- Pouget, A., Dayan, P. & Zemel, R. Inference and computation with population codes. *Annu. Rev. Neurosci.* **26**, 381–410 (2004).
- Meister, M., Wong, R. O. L., Baylor, D. A. & Shatz, C. J. Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science* **252**, 939–943 (1991).
- Arieli, A., Shoham, D., Hildesheim, R. & Grinvald, A. Coherent spatiotemporal patterns of ongoing activity revealed by real-time optical imaging coupled with single-unit recording in the cat visual cortex. *J. Neurophysiol.* **73**, 2072–2093 (1995).
- Weliky, M. & Katz, L. C. Correlational structure of spontaneous neuronal activity in the developing lateral geniculate nucleus in vivo. *Science* **285**, 599–604 (1999).
- Tsodyks, M., Kenet, T., Grinvald, A. & Arieli, A. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* **286**, 1943–1946 (1999).
- Chiu, C. & Weliky, M. Spontaneous activity in developing ferret visual cortex in vivo. *J. Neurosci.* **21**, 8906–8914 (2001).
- Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A. & Arieli, A. Spontaneously emerging cortical representations of visual attributes. *Nature* **425**, 954–956 (2003).
- Gao, W., Newman, D. E., Wormington, A. B. & Pallas, S. Development of inhibitory circuitry in visual and auditory cortex of postnatal ferrets: Immunocytochemical localization of GABAergic neurons. *J. Comp. Neurol.* **409**, 261–273 (1999).
- Anderson, J., Lampl, I., Reichova, I., Carandini, M. & Ferster, D. Stimulus dependence of two-state fluctuations of membrane potential in cat visual cortex. *Nature Neurosci.* **3**, 617–621 (2000).
- Cossart, R., Aronov, D. & Yuste, R. Attractor dynamics of network UP states in the neocortex. *Nature* **423**, 283–288 (2003).
- Shu, Y. S., Hasenstaub, A., Badoual, M., Bal, T. & McCormick, D. A. Barrages of synaptic activity control the gain and sensitivity of cortical neurons. *J. Neurosci.* **23**, 10388–10401 (2003).
- Gallant, J. L., Connor, C. E. & Van Essen, D. C. Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport* **9**, 1673–1678 (1998).
- Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
- Gawne, T. J. & Richmond, B. J. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* **13**, 2758–2771 (1993).
- Pol, G., Thiele, A., Hoffmann, K. P. & Panzeri, S. An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network Comput. Neural Syst.* **14**, 35–60 (2003).
- Nirenberg, S. & Latham, P. E. Decoding neuronal spike trains: How important are correlations? *Proc Natl Acad. Sci. USA* **100**, 1045–1050 (2003).
- Kara, P., Reinagel, P. & Reid, R. C. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron* **27**, 635–646 (2000).
- Hirsch, J. A. et al. Synaptic physiology of the flow of information in the cat's visual cortex in vivo. *J. Physiol. (Lond.)* **540**, 335–350 (2002).
- Attwell, D. & Laughlin, S. B. An energy budget for signaling in the grey matter of the brain. *J. Cereb. Blood Flow Metab.* **21**, 1133–1145 (2001).
- Lennie, P. The cost of cortical computation. *Curr. Biol.* **13**, 493–497 (2003).
- Weliky, M., Fiser, J., Hunt, R. H. & Wagner, D. N. Coding of natural scenes in primary visual cortex. *Neuron* **37**, 703–718 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by NIH (NEI) and the McKnight Foundation. We thank David Wagner for technical assistance. We also thank R. Aslin, D. Knill, D. Lee and K. Nordeen for comments. We also thank E. Romanski for supplying the ISCAN equipment.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.W. (weliky@cvs.rochester.edu).

A transmembrane protein required for acetylcholine receptor clustering in *Caenorhabditis elegans*

Christelle Gally¹, Stefan Eimer¹, Janet E. Richmond² & Jean-Louis Bessereau¹

¹INSERM U.497, École Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France

²University of Illinois at Chicago, Chicago, Illinois 60607, USA

Clustering neurotransmitter receptors at the synapse is crucial for efficient neurotransmission. Here we identify a *Caenorhabditis elegans* locus, *lev-10*, required for postsynaptic aggregation of ionotropic acetylcholine receptors (AChRs). *lev-10* mutants were identified on the basis of weak resistance to the anthelmintic drug levamisole, a nematode-specific cholinergic agonist that activates AChRs present at neuromuscular junctions (NMJs) resulting in muscle hypercontraction and death at high concentrations^{1–3}. In *lev-10* mutants, the density of levamisole-sensitive AChRs at NMJs is markedly reduced, yet the number of functional AChRs present at the muscle cell surface remains unchanged. LEV-10 is a transmembrane protein localized to cholinergic NMJs and required in body-wall muscles for AChR clustering. We also show that the LEV-10 extracellular region, containing five predicted CUB domains and one LDL domain, is sufficient to rescue AChR aggregation in *lev-10* mutants. This suggests a mechanism for AChR clustering that relies on extracellular protein–protein interactions. Such a mechanism is likely to be evolutionarily conserved because CUB/LDL transmembrane proteins similar to LEV-10, but lacking any assigned function, are expressed in the mammalian nervous system and might be used to cluster ionotropic receptors in vertebrates.

Genetic screens for *C. elegans* mutants that exhibit strong resistance to levamisole have identified four genes encoding AChR subunits and two genes that are required for the biosynthesis of levamisole-sensitive AChRs^{1–3}. However, no genes required for AChR clustering were cloned despite the large size of these screens. We hypothesized that impairing the function of such genes would generate subtle phenotypes for two reasons. First, unclustered levamisole-sensitive AChRs might remain functional if properly inserted into the plasma membrane, thus conferring levamisole-sensitivity. Second, there is an additional class of AChRs present at *C. elegans* NMJs that are activated by acetylcholine and nicotine but are insensitive to levamisole⁴. These receptors, of as yet unknown composition, might compensate for a decrease in levamisole-sensitive AChRs at the synapse.

We therefore performed a screen to isolate mutants that exhibited only weak resistance to levamisole. To facilitate the identification of mutated genes we used an insertional mutagenesis based on germline mobilization of the *Drosophila* transposon *Mos1* (ref. 5). We isolated a mutant allele, *kr26*, that resulted from a *Mos1* insertion whose interpolated genetic position was in the vicinity of the *lev-10* locus. A single mutant allele of *lev-10*, *x17*, was isolated previously in a levamisole-resistance screen but was not characterized at the molecular level¹. Using a genetic complementation test, we showed that *x17* and *kr26* are two alleles of the same gene, *lev-10*. Both *lev-10* mutants displayed a slight resistance to levamisole, when assayed by dose–response (Fig. 1a), but after 1 h of exposure to 1 mM levamisole, 100% of the *lev-10* mutants became paralysed. Although, in contrast to wild-type animals, *lev-10* mutants were able to survive while remaining hypercontracted at this elevated drug concentration. In addition, both *lev-10* mutants displayed marginal locomotory defects on plates. When movement was analysed in liquid medium, a subtle but significant movement impairment was

detected (Fig. 1b). These phenotypes suggested that mutating *lev-10* only partly impairs the function of levamisole-sensitive AChRs.

To analyse the expression of these receptors, we raised antibodies against UNC-29, a non- α -subunit of the levamisole-sensitive AChR in muscle⁶. In wild-type animals, UNC-29 was clustered along the ventral and dorsal cords and in the nerve ring where head muscles are innervated (Fig. 2a, c). In *lev-10* mutants, no detectable UNC-29 staining was observed along the ventral and dorsal nerve cords, and only weak staining remained in the nerve ring (Fig. 2b, d). To test whether the loss of UNC-29 clusters in *lev-10* mutants was due to the absence of cholinergic innervation, we immunostained cholinergic varicosities with an antibody against UNC-17, the vesicular acetylcholine transporter in *C. elegans*⁷. Staining patterns

were similar in wild-type (Fig. 2e) and *lev-10* mutant animals (Fig. 2f). In addition to cholinergic innervation, body-wall muscles are also innervated by GABAergic motoneurons. To determine whether *lev-10* was globally required for the formation of receptor aggregates or was specifically acting at cholinergic neuromuscular synapses, we immunostained the muscle GABA_A receptor UNC-49 (refs 8, 9). In both the wild type and *lev-10* mutants, GABA receptors were clustered along the nerve cords (Fig. 2g, h). The inability to detect AChRs by immunostaining could result from decreased receptor expression in *lev-10* mutants. Alternatively, a diffuse distribution of a wild-type number of receptors could be below our detection threshold. AChR expression was therefore assessed by western blot analysis of fractionated worm extracts (Fig. 2i). In *lev-10(kr26)* and *lev-10(x17)* extracts, the UNC-29 concentrations were similar to that in the wild type ($90 \pm 12\%$ ($n = 4$) and $123 \pm 17\%$ ($n = 4$), respectively), suggesting that AChR expression is not reduced in *lev-10* mutants.

To test whether the UNC-29 protein detected in *lev-10* mutants was assembled into functional receptors present at the muscle cell surface, we used electrophysiology⁴. Pressure-ejection of levamisole onto voltage-clamped body-wall muscles elicited similar currents in the wild type and in *lev-10* mutants (Fig. 3a, b), whereas in a *unc-29* *lev-10* double mutant no levamisole current was detected (data not shown). These data indicate that the overall expression level of functional levamisole-sensitive AChRs in *lev-10* mutants is comparable to that in the wild type. To analyse the synaptic population of levamisole-sensitive AChRs, we stimulated motoneurons in the ventral cord and recorded evoked currents in individual muscle cells. To eliminate currents due to activation of the GABA receptor UNC-49, we performed our analysis in an *unc-49(e407)* null mutant background. Furthermore, we used the nicotinic antagonist dihydro- β -erythroidine (DH β E) to block levamisole-insensitive AChRs present at NMJs⁴. In *lev-10;unc-49* double mutants, the size of the evoked response in the presence of DH β E was decreased by 77% compared with that of *unc-49* (Fig. 3c, d). The remaining evoked current was due to the activation of levamisole-sensitive AChRs, because *unc-29;unc-49* double mutants, which no longer

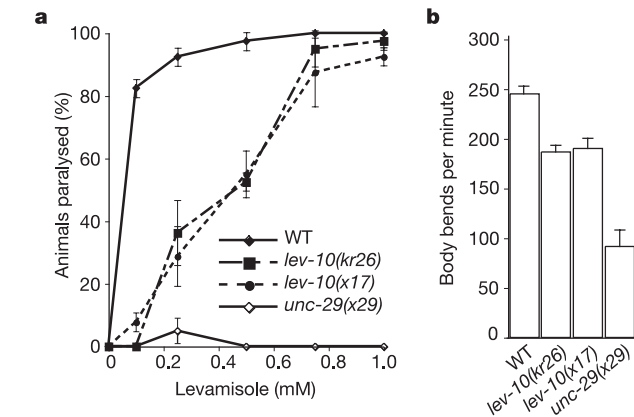


Figure 1 Phenotypic characterization of *lev-10* mutants. **a**, The levamisole dose-response curve indicates that *lev-10* mutants are only weakly resistant to levamisole when compared with *unc-29(x29)* mutants, which lack levamisole-sensitive AChRs. Error bars represent s.e.m. ($n = 4$ independent experiments). WT, wild type. **b**, *lev-10* mutants exhibit weak locomotory defects compared with the wild type in a thrashing assay (ANOVA test; $p < 0.01$) but are not as impaired as *unc-29(x29)* mutants ($p < 0.01$). Error bars represent s.e.m. ($n = 6$).

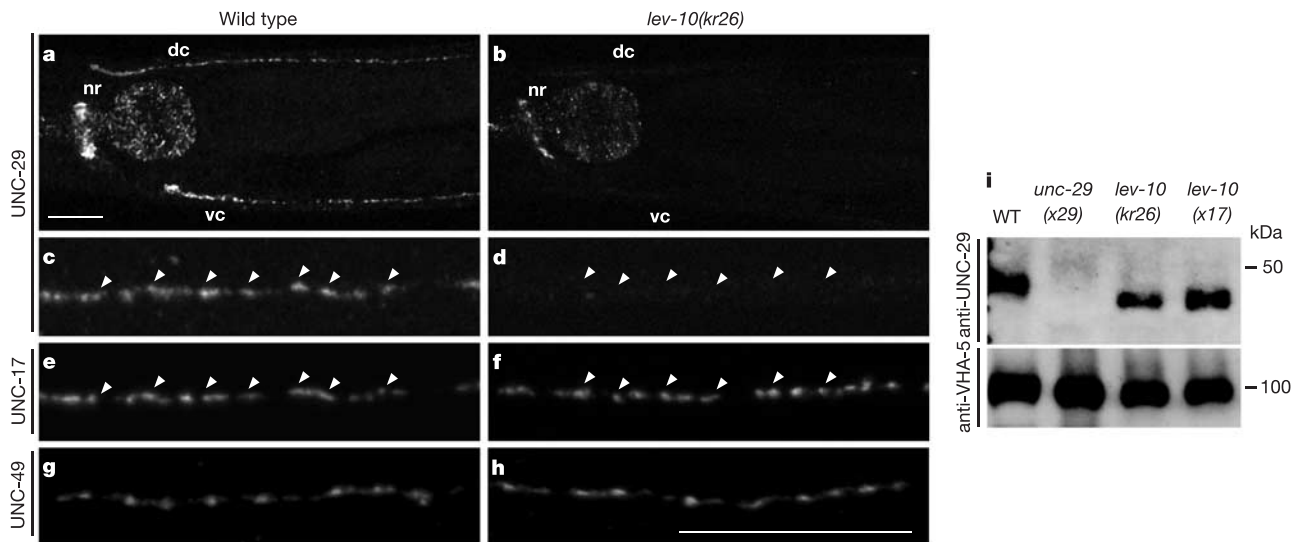


Figure 2 Mutation of *lev-10* results in the specific loss of levamisole-sensitive AChR clusters at neuromuscular junctions. **a-d**, UNC-29 localization detected by immunofluorescence with anti-UNC-29 antibodies. **a**, Shown are the nerve ring (nr) and the dorsal (dc) and ventral (vc) nerve cords in wild-type animals. **c**, Individual UNC-29 puncta at high magnification in the dorsal cord from the wild type. **b, d**, UNC-29 staining in *lev-10(kr26)* animals at magnifications as in **a** and **c**, respectively. The staining in the pharynx is non-specific (data not shown). **e**, Visualization of cholinergic varicosities by

co-immunostaining of the vesicular ACh transporter UNC-17 in wild-type animals shows that UNC-29 clusters are juxtaposed to cholinergic release sites (arrowheads). **f**, UNC-17 staining in *lev-10(kr26)* mutants. **g, h**, Immunostaining of the GABA receptor UNC-49 in wild-type animals (**g**) and *lev-10(kr26)* mutants (**h**). Scale bars, 20 μ m. **i**, Western blot with anti-UNC-29 and anti-VHA-5 antibodies on membrane fractions of *C. elegans* extracts. The UNC-29 protein has an apparent molecular mass of 47 kDa. VHA-5 detection is used for normalization. WT, wild type.

express levamisole-sensitive AChRs⁴, exhibited no evoked response in the presence of DH β E. In addition, the time to peak and decay time of the evoked current were increased in *lev-10;unc-49* compared with those in *unc-49* (6.44 ± 0.41 ms ($n = 7$) versus 4.64 ± 0.18 ms ($n = 7$), $p < 0.0017$, and 16.6 ± 6.9 ms ($n = 6$) versus 8.4 ± 0.43 ms ($n = 7$), $p < 0.0002$, respectively). These

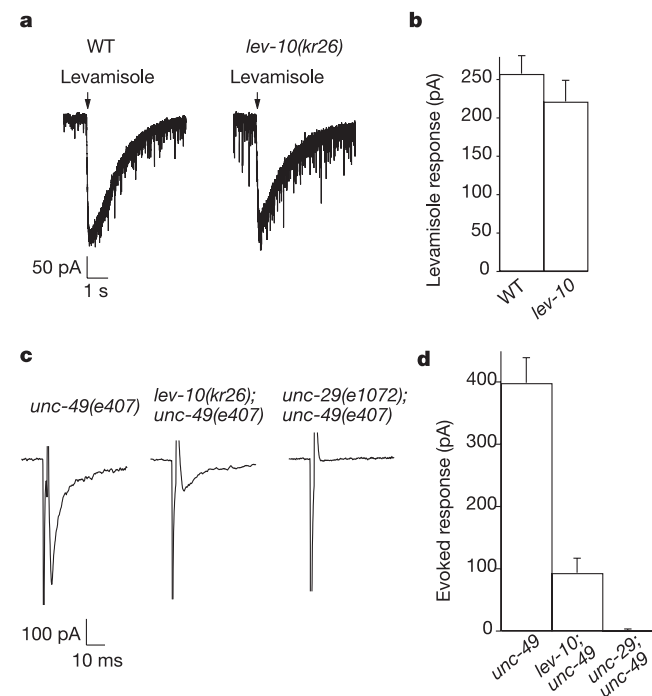


Figure 3 Levamisole-sensitive AChRs are functional but diffusely distributed in *lev-10* body-wall muscle. **a**, Currents recorded from voltage-clamped body wall muscles in response to pressure-ejection of levamisole (300 μ M) in wild-type (WT) and *lev-10(kr26)* mutants. **b**, Average amplitude of levamisole-elicited current. **c**, Evoked currents recorded in a body-wall muscle after eliciting neurotransmitter release by ventral nerve cord depolarization. Experiments were performed in an *unc-49(e407)* background to eliminate currents due to GABA receptor activation and in the presence of 5 μ M DH β E, which blocks the levamisole-insensitive AChRs. **d**, Average amplitude of evoked response. Error bars in **b** and **d** represent s.e.m.

kinetic alterations are consistent with a decreased ratio of synaptic versus perisynaptic receptors being activated by synaptic release of acetylcholine in the *lev-10* background. Analysis of evoked response amplitudes in the absence of DH β E in *lev-10;unc-49* and *unc-49* mutants did not reveal any significant difference (3329 ± 221 pA ($n = 6$) versus 2904 ± 291 pA ($n = 7$), respectively), thus suggesting that the expression and localization of the levamisole-insensitive AChRs present at the NMJ were not affected in *lev-10* mutants. In combination with the immunostaining data, these results indicate that *lev-10* is required specifically for the clustering of levamisole-sensitive AChRs at the synapse.

We cloned *lev-10* using inverse polymerase chain reaction (PCR) to identify the genomic position of the *kr26::Mos1* insertion (Fig. 4a) and confirmed its identity by rescue experiments (Supplementary Table S1). Interestingly, *lev-10* overlaps with *eat-18*, a gene required for the function of AChRs in pharyngeal muscle^{10,11}. Mutation of *eat-18* does not confer levamisole resistance. We confirmed that these genes are distinct by genetic complementation (data not shown) and by rescuing *lev-10* mutants with a genomic fragment carrying the *eat-18(ad1110)* nonsense mutation (Supplementary Table S1). *lev-10* is predicted to encode a type I transmembrane protein (Fig. 4b). The extracellular part of the protein contains five predicted CUB domains and one LDLa domain. These domains are present in a wide variety of secreted and membrane-bound proteins and mediate protein–protein interactions (for reviews see refs 12–14). Alternative splicing of *lev-10* generates two transcripts, *lev-10a* and *lev-10b* (Fig. 4b). The *lev-10b* splice variant represents less than 10% of *lev-10* mRNAs (data not shown) and codes for a LEV-10 isoform that differs from LEV-10A in the transmembrane region and contains virtually no intracellular region.

In wild-type animals, LEV-10 is concentrated at cholinergic NMJs (Fig. 5a, c, e). Double labelling experiments with an antibody against the vesicular acetylcholine transporter UNC-17 (ref. 15) demonstrated that $93 \pm 3\%$ of LEV-10 puncta were associated with cholinergic varicosities (mean $\pm$ SEM, 104 puncta counted in seven worms). However, three-dimensional analysis of confocal image stacks revealed that LEV-10 staining was juxtaposed to, but did not overlap, UNC-17 distribution (Fig. 5f). To determine whether LEV-10 functions postsynaptically, we expressed LEV-10A or LEV-10B under the control of the muscle-specific promoter *myo-3* (ref. 16) in *lev-10(kr26)* animals. Both proteins rescued behavioural defects and UNC-29 synaptic clustering when expressed in muscle (Supplementary Table S1). Genetic mosaic

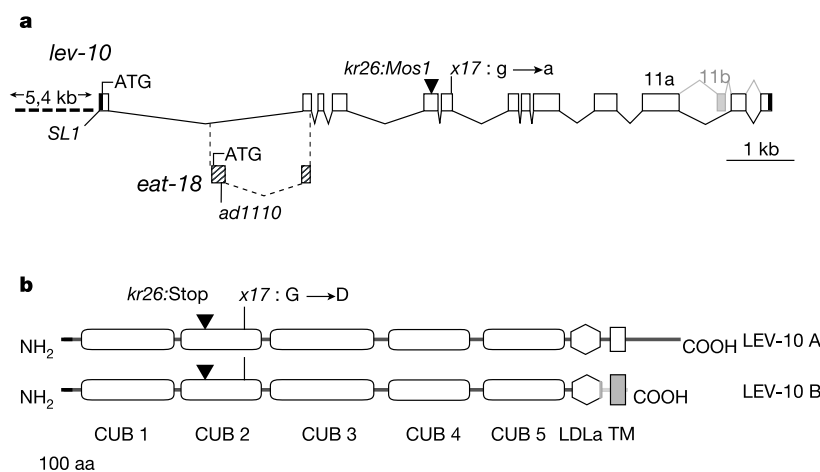


Figure 4 *lev-10* encodes a CUB domain-rich transmembrane protein. **a**, Genomic organization of *lev-10*. Open boxes, coding regions; black boxes, 5' and 3' untranslated region; ATG, translational start site; SL1, SL1 trans-spliced leader. The first intron of *lev-10* contains the first exon of the gene *eat-18* (hatched boxes). The *eat-18* exon is spliced to the second exon of *lev-10* by using a different frame, which ends 16 bp after the splice

site. *ad1110*, nonsense mutation in the first exon of *eat-18*. **b**, Predicted structure of the LEV-10 isoforms. Horizontal black line, signal peptide; CUB, complement, urchin epidermal growth factor, and bone morphogenetic protein domain; LDLa, low-density lipoprotein receptor domain class A; TM, transmembrane region; aa, amino acids. Domain predictions were based on SMART (<http://smart.embl-heidelberg.de>).

analysis (Supplementary Information) confirmed that LEV-10 is required cell autonomously in postsynaptic muscle cells for AChR clustering at NMJs.

Recent results have suggested that proteins involved in neurotransmitter receptor clustering do not accumulate at the synapse in the absence of the receptors¹⁷. To test this possibility in our system, we analysed the distribution of LEV-10 in animals lacking levamisole-sensitive AChRs. In *unc-29(x29)* and *unc-38(x20)* AChR subunit mutants, no LEV-10 was detected in ventral and dorsal nerve cords by immunofluorescence (Fig. 5b and data not shown) even though presynaptic cholinergic varicosities differentiated normally (Fig. 5d). In parallel, LEV-10 expression level was assessed by western blot analysis (Fig. 5g). LEV-10 was detected as a 120-kDa protein present in the membrane fraction of wild-type worm extracts. This band was absent from *lev-10(kr26)* extracts and was markedly reduced in *lev-10(x17)*. In *unc-29* and *unc-38* extracts, LEV-10 concentrations were decreased only slightly in comparison with those in the wild type ($72\% \pm 11$ ($n = 4$) and $81\% \pm 4$ ($n = 5$), respectively), indicating that a lack of levamisole-sensitive AChRs does not alter LEV-10 expression level. Because LEV-10 is expressed but fails to accumulate at synapses in the absence of levamisole-sensitive AChRs, we cannot exclude the possibility that LEV-10 requires AChRs to reach the plasma membrane, although no intracellular staining of LEV-10 is seen by immunofluorescence in *unc-29* and *unc-38* mutants. Alternatively, LEV-10 might interact directly or indirectly with AChRs in a complex that is recruited or stabilized at the synapse.

Because most characterized ionotropic receptor clustering proteins are cytoplasmic, relevant interactions are thought to occur on the cytoplasmic side of the postsynaptic membrane^{18,19}. However, complexes formed on the extracellular side of the postsynaptic membrane might also be critical^{20–23}. To test this possibility, we fused the extracellular part of LEV-10 to the human CD4 transmembrane domain. Expression of this chimaeric protein in muscle rescued the defects in levamisole sensitivity, locomotion and AChR clustering of *lev-10(kr26)* animals. Furthermore, we overexpressed a green fluorescent protein (GFP)-tagged version of LEV-10 truncated before the transmembrane segment. This protein was secreted from muscle cells into the pseudocoelomic cavity (data not shown) but was still able to rescue *lev-10(kr26)* mutant phenotypes (Supplementary Table S1). The function of LEV-10 in AChR clustering therefore seems to involve only extracellular interactions.

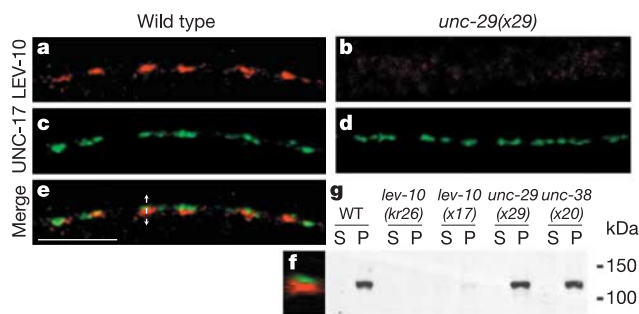


Figure 5 LEV-10 is a synaptic protein that requires levamisole-sensitive AChRs for proper localization but not for expression. **a**, LEV-10A immunostaining in the dorsal cord of a wild-type animal. **c**, UNC-17 immunostaining of the same animal labels cholinergic varicosities. **e**, Merged images. **f**, Z-optical projection through the entire stack of confocal images at the level of the dashed arrow in **e**. **b**, **d**, LEV-10A (**b**) and UNC-17 (**d**) immunostaining in the dorsal cord of an *unc-29(x29)* mutant. Scale bar, 10 μ m. **g**, Western blot analysis of fractionated *C. elegans* extracts. P, pellet; S, cytosolic supernatant. The LEV-10 transmembrane protein has an apparent molecular mass of about 120 kDa. Mutants of the levamisole-sensitive AChR subunits *unc-29(x29)* and *unc-38(x20)* have LEV-10 concentrations at the membrane comparable to those of the wild type.

LEV-10 is the first example of a CUB/LDL protein involved in the synaptic clustering of AChRs. The presence of multiple predicted protein–protein interaction domains in the extracellular region indicates that LEV-10 might bind multiple partners. Because we have so far been unable to demonstrate direct interactions between LEV-10 and AChRs, LEV-10 might be indirectly involved in the recruitment of signalling molecules that, in turn, cause AChR clustering. However, the interdependence between LEV-10 and AChR for synaptic localization is consistent with a model that would involve a set of interactions between LEV-10, AChRs or AChR-associated proteins, and a synaptic determinant used to nucleate clustering. Along this line, another *C. elegans* CUB-domain-rich transmembrane protein, SOL-1, has recently been shown to physically interact with glutamate receptors and to be required for glutamate-gated currents through an as yet unidentified mechanism²⁴.

Of 145 CUB-domain-containing mouse proteins present in non-redundant databases, the first two CUB domains of LEV-10 are most similar to those present in NETO2 (ref. 25) (26% identity, 43% similarity). NETO2 and its paralogue NETO1/BTCL1 (refs 25, 26) are predicted type I transmembrane proteins containing two CUB domains and one LDLA domain in their extracellular region. The two *NETO* genes are specifically expressed in retina and brain, but their function is unknown. It is therefore tempting to speculate that LEV-10 and NETO proteins are members of a novel class of membrane-spanning proteins engaged in postsynaptic domain organization by means of extracytoplasmic interactions at the synapse. □

Methods

Cloning of *lev-10*

N2 worms were mutagenized by germline mobilization of the *Drosophila* transposon *Mos1* (ref. 5). Young-adult F2 worms were screened for resistance to 1 mM levamisole 3–5 h after transfer to drug-containing plates. In EN 26 [*lev-10(kr26::Mos1)*], a *Mos1* insertion was localized in a predicted exon of the open reading frame Y105E8A.7a, at position 14,403,250 of chromosome I by using inverse PCR (WormBase, www.wormbase.org).

Rescue experiments were performed with a genomic fragment covering the Y105E8A.7a coding region plus 5 kilobases (kb) upstream of the translational start site and 0.21 kb downstream of the *lev-10a* stop codon. This 15-kb fragment was amplified from N2 or *eat-18(ad1110)*¹¹ genomic DNA and was injected at $0.85 \text{ ng } \mu\text{l}^{-1}$ with the use of pTG96 (*sur-5::GFP*)²⁷ as a co-injection marker at $100 \text{ ng } \mu\text{l}^{-1}$. Rescue was scored on the basis of survival on 1 mM levamisole for 16 h.

Tissue-specific rescue

lev-10 complementary DNAs were cloned by PCR after reverse transcription, and sequenced. *SL1* splicing was established by PCR with an SL1 primer and a primer in *lev-10* exon 13, and by sequencing the expressed sequence tag yk796a04.3'. PCR with rapid amplification of cDNA ends (RACE-PCR) was used to localize the polyadenylation site. *lev-10a* and *lev-10b* cDNAs were subcloned into pPD115.62 under the control of the *myo-3* promoter. The CD4 transmembrane domain amplified from the human CD4 cDNA (GenBank accession number M12807) was subcloned in frame into *Pmyo-3::lev-10a* and a stop codon was introduced immediately after the CD4 transmembrane domain. The secreted *gfp-lev-10s* was obtained by removing the CD4 transmembrane domain from *Pmyo-3::lev-10-CD4* and inserting a GFP cDNA immediately after the *lev-10* signal peptide. All constructs were injected at $20 \text{ ng } \mu\text{l}^{-1}$ together with pTG96 ($80 \text{ ng } \mu\text{l}^{-1}$) except for *Pmyo-3::gfp-lev-10s* ($10 \text{ ng } \mu\text{l}^{-1}$).

Levamisole dose–response curve

Young adult worms were scored blind for paralysis after 1 h exposure to levamisole. A distance of one body length of forward movement after mechanical stimulus was required to score a worm as non-paralysed.

Electrophysiological studies

Electrophysiological methods were performed as described previously⁴. Muscle recordings were made in the whole-cell voltage-clamp configuration (holding potential -60 mV) with an EPC-10 patch-clamp amplifier and digitized at 1 kHz. Data were acquired by Pulse software (HEKA). The bath solution contained 150 mM NaCl, 5 mM KCl, 5 mM CaCl_2 , 1 mM MgCl_2 , 10 mM glucose and 15 mM HEPES, pH 7.35, about 340 mOsm. The pipette solution was prepared as described previously⁴. Subsequent analysis and graphing were performed with Pulsefit (HEKA) and Igor Pro. All statistically derived values are given as means $\pm$ s.e.m.

Antibody production and immunocytochemistry

UNC-29: a DNA fragment encoding UNC-29 amino acids 348–431 was inserted into pGEX-3X (Amersham Biosciences). The glutathione-S-transferase (GST)–UNC-29

fusion protein was expressed in *Escherichia coli* and purified in accordance with the manufacturer's protocol. Rabbits were injected with 100 µg of fusion protein and boosted three times with 100 µg each.

LEV-10: two synthetic peptides (Eurogentec) corresponding to the LEV-10A amino acids 847–861 and 892–906 were injected into rabbits as described for UNC-29. Both antibodies were purified as described previously²⁸ by using the fusion proteins GST–UNC-29 or GST–LEV-10A (amino acids 836–906 in pGEX-3X) blotted on nitrocellulose. Immunostaining was performed as described⁹. UNC-29 antibody was used at a dilution of 1:250, and LEV-10 antibody at 1:300. For double-labelling experiments, UNC-17 monoclonal antibody¹⁵ was diluted at 1:500 and incubated for 1 h; after 1 h of washing, UNC-29 or LEV-10 antibodies were incubated overnight. The secondary antibody, Cy3-labelled goat anti-rabbit IgG (H + L) (Jackson ImmunoResearch Laboratories), was diluted at 1:900 and the secondary antibody, Alexa488-labelled goat anti-mouse (Molecular Probes) at 1:200.

Protein extraction and western blotting

A mixed staged population of worms (500 µl) was frozen at –80 °C until use. For extraction, worm pellets were ground under liquid nitrogen and thawed on ice. While thawing, 6–10 volumes of ice-cold homogenization buffer (20 mM HEPES pH 7.4, 10 mM KCl, 1 mM EDTA, 400 µM Pefabloc (Roche) and Complete Mini Protease inhibitor cocktail (Roche)) were added and the suspension was further homogenized with ten strokes with the use of a 2-ml tight-fitting glass tissue homogenizer. Afterwards an equal volume of homogenization buffer containing 0.5 M sucrose was added and the suspension was centrifuged twice at 2,000g for 10 min to remove worm debris. The resulting nuclear pellets were pooled and extracted twice with 5 ml of homogenization buffer containing 0.25 M sucrose. The post-nuclear supernatants were pooled and subsequently centrifuged at 150,000g for 1 h. Equal amounts (about 30 µg) of the resulting cytosolic supernatant and membrane pellet were separated by SDS–polyacrylamide-gel electrophoresis and blotted onto nitrocellulose membranes. The membranes were subsequently probed with purified anti-LEV-10 serum (dilution 1:1000), anti-UNC-29 (1:600) or anti-VHA-5 (1:3000) (M. Labouesse, unpublished observations) and horseradish-peroxidase-conjugated goat anti-rabbit antibodies (DAKO) and revealed with LumiLight reagents (Roche).

Received 3 June; accepted 30 July 2004; doi:10.1038/nature02893.

- Lewis, J. A., Wu, C. H., Berg, H. & Levine, J. H. The genetics of levamisole resistance in the nematode *Caenorhabditis elegans*. *Genetics* **95**, 905–928 (1980).
- Lewis, J. A., Wu, C. H., Levine, J. H. & Berg, H. Levamisole-resistant mutants of the nematode *Caenorhabditis elegans* appear to lack pharmacological acetylcholine receptors. *Neuroscience* **5**, 967–989 (1980).
- Lewis, J. A. *et al.* Cholinergic receptor mutants of the nematode *Caenorhabditis elegans*. *J. Neurosci.* **7**, 3059–3071 (1987).
- Richmond, J. E. & Jorgensen, E. M. One GABA and two acetylcholine receptors function at the *C. elegans* neuromuscular junction. *Nature Neurosci.* **2**, 791–797 (1999).
- Bessereau, J. L. *et al.* Mobilization of a *Drosophila* transposon in the *Caenorhabditis elegans* germ line. *Nature* **413**, 70–74 (2001).
- Fleming, J. T. *et al.* *Caenorhabditis elegans* levamisole resistance genes *lev-1*, *unc-29*, and *unc-38* encode functional nicotinic acetylcholine receptor subunits. *J. Neurosci.* **17**, 5843–5857 (1997).
- Alfonso, A., Grundahl, K., Duerr, J. S., Han, H. P. & Rand, J. B. The *Caenorhabditis elegans* *unc-17* gene: a putative vesicular acetylcholine transporter. *Science* **261**, 617–619 (1993).
- Bamber, B. A., Beg, A. A., Twyman, R. E. & Jorgensen, E. M. The *Caenorhabditis elegans* *unc-49* locus encodes multiple subunits of a heteromultimeric GABA receptor. *J. Neurosci.* **19**, 5348–5359 (1999).
- Gally, C. & Bessereau, J. L. GABA is dispensable for the formation of junctional GABA receptor clusters in *Caenorhabditis elegans*. *J. Neurosci.* **23**, 2591–2599 (2003).
- Raizen, D. M., Lee, R. Y. & Avery, L. Interacting genes required for pharyngeal excitation by motor neuron MC in *Caenorhabditis elegans*. *Genetics* **141**, 1365–1382 (1995).
- McKay, J. P., Raizen, D. M., Gottschalk, A., Schafer, W. R. & Avery, L. *eat-2* and *eat-18* are required for nicotinic neurotransmission in the *C. elegans* pharynx. *Genetics* **166**, 161–169 (2004).
- Bork, P. & Beckmann, G. The CUB domain. A widespread module in developmentally regulated proteins. *J. Mol. Biol.* **231**, 539–545 (1993).
- Christensen, E. I. & Birn, H. Megalin and cubilin: multifunctional endocytic receptors. *Nature Rev. Mol. Cell Biol.* **3**, 256–266 (2002).
- Herz, J. & Bock, H. H. Lipoprotein receptors in the nervous system. *Annu. Rev. Biochem.* **71**, 405–434 (2002).
- Duerr, J. S., Gaskin, J. & Rand, J. B. Identified neurons in *C. elegans* coexpress vesicular transporters for acetylcholine and monoamines. *Am. J. Physiol. Cell Physiol.* **280**, C1616–C1622 (2001).
- Okkema, P. G., Harrison, S. W., Plunger, V., Aryana, A. & Fire, A. Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* **135**, 385–404 (1993).
- Ono, F., Mandel, G. & Brehm, P. Acetylcholine receptors direct rapsyn clusters to the neuromuscular synapse in zebrafish. *J. Neurosci.* **24**, 5475–5481 (2004).
- Bredt, D. S. & Nicoll, R. A. AMPA receptor trafficking at excitatory synapses. *Neuron* **40**, 361–379 (2003).
- Choquet, D. & Triller, A. The role of receptor diffusion in the organization of the postsynaptic membrane. *Nature Rev. Neurosci.* **4**, 251–265 (2003).
- O'Brien, R. J. *et al.* Synaptic clustering of AMPA receptors by the extracellular immediate-early gene product Narp. *Neuron* **23**, 309–323 (1999).
- Dalva, M. B. *et al.* EphB receptors interact with NMDA receptors and regulate excitatory synapse formation. *Cell* **103**, 945–956 (2000).
- Passafaro, M., Nakagawa, T., Sala, C. & Sheng, M. Induction of dendritic spines by an extracellular domain of AMPA receptor subunit GluR2. *Nature* **424**, 677–681 (2003).
- Tomita, S., Fukata, M., Nicoll, R. A. & Bredt, D. S. Dynamic interaction of stargazin-like TARPs with cycling AMPA receptors at synapses. *Science* **303**, 1508–1511 (2004).
- Zheng, Y., Mellem, J. E., Brockie, P. J., Madsen, D. M. & Maricq, A. V. SOL-1 is a CUB-domain protein required for GLR-1 glutamate receptor function in *C. elegans*. *Nature* **427**, 451–457 (2004).
- Stohr, H., Berger, C., Fröhlich, S. & Weber, B. H. A novel gene encoding a putative transmembrane protein with two extracellular CUB domains and a low-density lipoprotein class A module: isolation of alternatively spliced isoforms in retina and brain. *Gene* **286**, 223–231 (2002).

- Michishita, M. *et al.* A novel gene, Btcl1, encoding CUB and LDLA domains is expressed in restricted areas of mouse brain. *Biochem. Biophys. Res. Commun.* **306**, 680–686 (2003).
- Yochem, J., Gu, T. & Han, M. A new marker for mosaic analysis in *Caenorhabditis elegans* indicates a fusion between hyp6 and hyp7, two major components of the hypodermis. *Genetics* **149**, 1323–1334 (1998).
- Miller, K. G., Emerson, M. D., McManus, J. R. & Rand, J. B. RIC-8 (Synembryn): a novel conserved protein that is required for G_q signaling in the *C. elegans* nervous system. *Neuron* **27**, 289–299 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Lewis for the *lev-10(x17)* strain, M. Labouesse for the anti-VAH-5 antibodies, J. Rand for the anti-UNC-17 antibodies, M. Han for the pTG96 plasmid, Y. Kohara for the clone yk796a04, A. Fire for the GFP vectors, the *Caenorhabditis* Genetic Center for strains, R. Weimer for critical reading of the manuscript, and I. Nuez and H. Gendrot for technical help. C.G. was supported by a fellowship from the Ministère de la Recherche and by the Association pour la Recherche contre le Cancer. S.E. is an EMBO fellow. This work was funded by the Institut National de la Santé et de la Recherche Médicale and the Association Française contre les Myopathies. J.R. was supported by the NIH grant RO1NS41477-03.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.-L.B. (jlbesse@wotan.ens.fr). The EMBL database accession numbers for *lev-10a* and *lev-10b* cDNAs are BN000434 and BN000435, respectively.

The rice leaf blast pathogen undergoes developmental processes typical of root-infecting fungi

Ane Sesma & Anne E. Osbourn

The Sainsbury Laboratory, John Innes Center, Colney Lane, Norwich NR4 7UH, UK

Pathogens have evolved different strategies to overcome the various barriers that they encounter during infection of their hosts¹. The rice blast fungus *Magnaporthe grisea* causes one of the most damaging diseases of cultivated rice and has emerged as a paradigm system for investigation of foliar pathogenicity. This fungus undergoes a series of well-defined developmental steps during leaf infection, including the formation of elaborate penetration structures (appressoria). This process has been studied in great detail², and over thirty *M. grisea* genes that condition leaf infection have been identified³. Here we show a new facet of the *M. grisea* life cycle: this fungus can undergo a different (and previously uncharacterized) set of programmed developmental events that are typical of root-infecting pathogens. We also show that root colonization can lead to systemic invasion and the development of classical disease symptoms on the aerial parts of the plant. Gene-for-gene type specific disease resistance that is effective against rice blast in leaves also operates in roots. These findings have significant implications for fungal development, epidemiology, plant breeding and disease control.

Because rice is the staple food for half of the global population, rice blast is a constant threat to the world's food supply. Control strategies depend on use of resistant cultivars and application of fungicides, although neither of these methods is particularly effective⁴. The development of durable, environmentally friendly strategies for the control of rice blast disease will depend on a better understanding of the disease process. To this end, the sequence of the *M. grisea* genome has been completed with the objective of gaining an intimate knowledge of the pathogen and of factors governing disease⁴. Recent changes in fungal taxonomy have led to the reclassification of *M. grisea* into the newly established Magnaporthaceae family⁵. This family includes the soil-borne pathogen *Gaeumannomyces graminis*^{6,7}, which causes the take-all

disease of cereals, along with other close relatives that also infect roots, such as *M. poae*⁸ (the causal agent of summer patch disease of turfgrasses) and *M. rhizophila*⁹ (a wheat root-infecting fungus). We have recently reported that *M. grisea* can also cause lesions when inoculated onto cereal roots¹⁰, although we did not investigate the developmental processes associated with root infection. Here we show that this paradigm foliar pathogen has all the features of a classical root pathogen. These results have profound implications for the way that we think about plant disease, both from an evolutionary and an aetiological perspective.

M. grisea undergoes a series of well-defined developmental steps before it infects rice leaves². Conidia germinate on the leaf surface and the germ tubes differentiate into specialized penetration structures known as appressoria, which are heavily melanized. The appressoria then build up the tremendous turgor pressure required for mechanical penetration of the tough leaf surface^{11,12}. We used green fluorescent protein (GFP)-tagged *M. grisea* strains and also chlorazole black E staining to investigate the developmental events associated with root infection. Remarkably, we found that *M. grisea* can undergo a range of developmental processes that have not previously been described for this fungus and that are typical of root pathogens. The melanized appressoria associated with classical foliar infection (Fig. 1a) were not observed on the surface of rice roots when either conidia or mycelium were used as inoculum. In contrast, hyphal swellings resembling the simple penetration structures (hyphopodia) of root-infecting fungi were evident at infection sites, often with associated infection pegs (Fig. 1b, c). On barley there was more extensive surface colonization and dark runner hyphae were clearly visible on the root surface (Fig. 1d, e). These hyphae were very similar to those formed by the soil-borne take-all pathogen *G. graminis* (Fig. 1l). Following penetration, intra- and intercellular growth of the fungus was visible in the epidermal and cortical cell layers of the root. The intracellular hyphae were thick and bulbous with constrictions in places where they crossed the plant cell wall (Fig. 1f). This invasive growth phase is similar to that observed during foliar infection by *M. grisea*¹³. Other features typical of root-infecting fungi included microsclerotia, which

developed on the surface of the roots (Fig. 1g, h). Round brown structures were also seen on the root surface and inside epidermal and cortical cells (Fig. 1i). These resembled the growth cessation structures formed by *Gaeumannomyces* and related root-colonizing fungi⁶ (Fig. 1l). The above structures were all seen during colonization of rice roots by *M. grisea* strain Guy11 (ref. 14). In addition, aggregates of swollen hyphal cells with pores were evident on barley roots (Fig. 1j, k); these also bore striking similarity to structures formed by root-infecting relatives (Fig. 1m).

The ability to produce melanized appressoria is an absolute requirement for leaf infection¹⁵. We generated GFP-expressing transformants of the melanin-deficient *buf* mutant 1.R.22 (ref. 15) and the corresponding parental strain 4091.5.8 (ref. 16) to investigate the root infection process in more detail. 1.R.22 produces melanin-deficient appressoria and so is unable to penetrate barley leaves, whereas 4091.5.8 is fully pathogenic¹⁵. Both strains produce necrotic lesions on barley roots (Fig. 2a). Confocal microscopy confirmed that both the wild type and melanin-deficient strains are able to penetrate the root by means of simple hyphopodia-like structures (Fig. 2b). We also show that cyclic AMP (cAMP)-mediated signalling processes that are essential for formation of functional appressoria are not required for penetration of roots. *CPKA* encodes a catalytic subunit of a cAMP-dependent protein kinase required for appressorium morphogenesis^{3,17,18}. Thus, *cpkA* mutants (also in a 4091.5.8 background)¹⁸ were unable to form functional appressoria and were defective in penetration of leaf surfaces. However these mutants were not compromised in their ability to form hyphopodia and penetrate intact roots (Fig. 2a, b). This indicates that there are critical differences between the signal transduction pathway requirements for penetration of leaves and roots. Once inside the root *M. grisea* was highly invasive and progressed from the cortical cells through the endodermis and into the stele (Fig. 2c).

We then asked whether *M. grisea* shares common genetic requirements for root infection with other 'classical' root-infecting pathogens. Very little is known about fungal genes required for root colonization. However FOW1, a predicted mitochondrial carrier protein, has been shown to condition root infection in the wilt

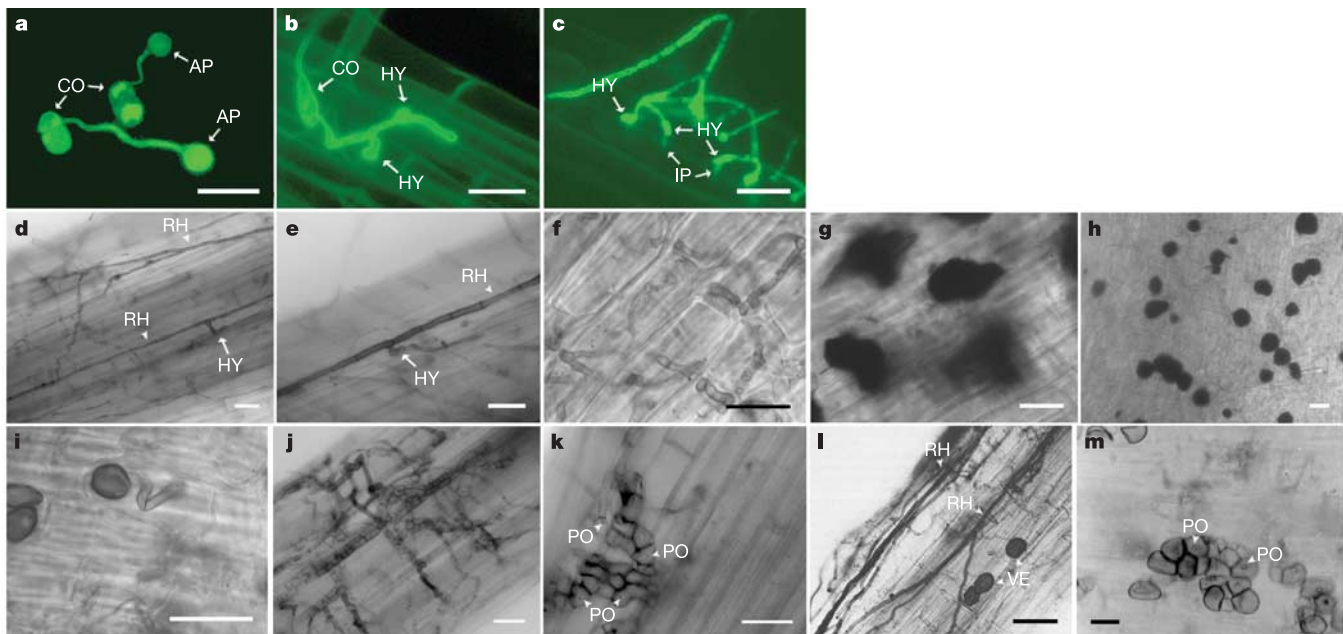


Figure 1 *M. grisea* undergoes developmental processes typical of root-infecting fungi. **a–c**, GFP-tagged *M. grisea* (Guy11) forms classical appressoria (AP) on a hydrophobic surface (**a**) and simple hyphopodia (HY) and infection pegs (IP) on rice roots (cultivar CO39) (**b, c**). **d–k**, Guy11-infected barley (**d, e, j, k**) and rice (**f–i**) roots stained with chlorazole black E showing: dark runner hyphae (RH) and simple hyphopodia (**d, e**); bulbous infection

hyphae invading epidermal cells (**f**); microsclerotia (previously reported in culture)³⁰ (**g, h**); vesicles (**i**), and swollen cells (**j, k**) with pores (PO). **l, m**, Structures produced by related root-infecting fungi (*Phialophora* spp.); vesicles (VE) and runner hyphae (RH) (**l**); swollen cells with pores (PO) (**m**); reproduced with permission from refs 6 and 7, respectively. Scale bar, 25 µm.

fungus *Fusarium oxysporum*¹⁹. We identified a *FOW1* homologue in the *M. grisea* genome sequence (Fig. 3a) and generated GFP-tagged gene replacement mutants (Fig. 3b). Deletion of *MgFOW1* caused a substantial reduction in root browning (Fig. 3c). Confocal microscopy showed a striking effect of *FOW1* deletion on root colonization. The Δ *MgFOW1* mutant, like the wild type, was able to colonize the surface of rice roots (Fig. 3d, left panels). However, growth of the mutant was arrested within the second cell layer, whereas the wild type strain progressed to the interior of the root (Fig. 3d, middle and right panels). The reduced root browning and the restricted root colonization phenotypes were both fully restored by complementation with the *MgFOW1* gene (data not shown). These experiments confirm that *M. grisea* and *F. oxysporum* share a common genetic requirement for root infection. The function of *FOW1*-like proteins is not known. *FOW1* shares close sequence similarity with a yeast protein (YMR241w) required for tricarboxylic acid transport¹⁹. However Δ *FOW1* mutants show normal saprophytic growth and YMR241w and *F. oxysporum* *FOW1* are not functionally interchangeable, leading to speculation that *FOW1* may have a specialized function in plants¹⁹. *MgFOW1* is less closely related to YMR241w than is *FOW1*. Δ *MgFOW1* mutants, like Δ *FOW1* mutants, are unimpaired in their ability to utilize glycerol as a carbon source (in contrast to Δ YMR241w mutants)¹⁹ (data not shown). Also, the deletion of *MgFOW1* has no effect on growth and conidiation on a range of rich and minimal media, indicating that *MgFOW1* is dispensable for saprophytic growth. Interestingly, Δ *MgFOW1* mutants are unaffected in their ability to form appressoria on plastic surfaces (data not shown) and in pathogenicity to leaves (Fig. 3c), and so the requirement of *MgFOW1* for plant colonization is tissue specific. Elucidation of the function of *FOW1*-like proteins represents an important future challenge.

Significantly, we have shown that *M. grisea* can spread from the roots of rice plants to the aerial tissues and cause disease. In infection studies with GFP-tagged *M. grisea* we found that up to 10% of the rice seedlings in our root infection experiments developed typical blast symptoms on the aerial parts of the plant. These included lesions on the leaves, collar rot and diamond-shaped necrotic lesions at the base of the stem (Fig. 4a–c). When the stems of infected seedlings were surface-sterilized and cut sections were placed on potato dextrose agar containing hygromycin, up to 50% of the stem sections yielded GFP-expressing colonies (data not shown). Confocal microscopy of diseased seedlings confirmed the presence of GFP-expressing hyphae within affected tissues and showed that the fungus can spread through the vascular tissue (Fig. 4c). We believe this to be the first report of systemic plant invasion by *M. grisea*. These data imply that root infection may contribute to the establishment of rice blast disease in the field, a possibility that to our knowledge has so far not been considered.

Because our experiments indicated that root colonization can lead to classical disease symptoms we next asked whether specific gene-for-gene type disease resistance mechanisms that are effective against rice blast above ground also function in roots. The rice cultivar CO39 contains the disease resistance gene *Pi-CO39(t)*, which confers specific resistance to *M. grisea* strains with the matching *AVR1-CO39* avirulence gene¹⁶. We constructed GFP-expressing transformants of a *M. grisea* Guy11 strain containing the avirulence gene *AVR1-CO39* (ref. 20) in order to establish whether *Pi-CO39(t)*-mediated disease resistance is effective in roots. GFP-expressing transformants were first confirmed to be avirulent on leaves of rice cultivar CO39 and fully virulent on the susceptible cultivar Nipponbare (data not shown). They were then tested on roots. The Guy11 control strain and the transformants expressing *AVR1-CO39* caused comparable levels of necrosis on roots of the two rice cultivars Nipponbare and CO39 (data not shown). Guy11 proliferated in the roots of rice cultivar CO39 (Fig. 4d) as expected. The Guy11 transformants expressing *AVR1-CO39* were able to proliferate within the stele of roots of

rice cultivar Nipponbare, which lacks a functional *Pi-CO39(t)* (Fig. 4e), but were only able to colonize the outermost cell layer of roots of rice cultivar CO39 (Fig. 4f). These data indicate that *Pi-CO39(t)*-mediated disease resistance is effective in rice roots. The demonstration that specific disease resistance is maintained in roots is a further indication of the biological significance of the root infection process. These findings also suggest that there may be a need for maintenance of rice blast resistance in roots, as redundant disease resistance functions can impose a fitness cost²¹.

Plant diseases are generally characterized by the tissue types that they affect. This may be misleading when considering disease epidemiology. Here we have used cytological and genetic analysis to show that *M. grisea* is capable of undergoing a range of developmental processes that are clearly distinct from the well-defined set

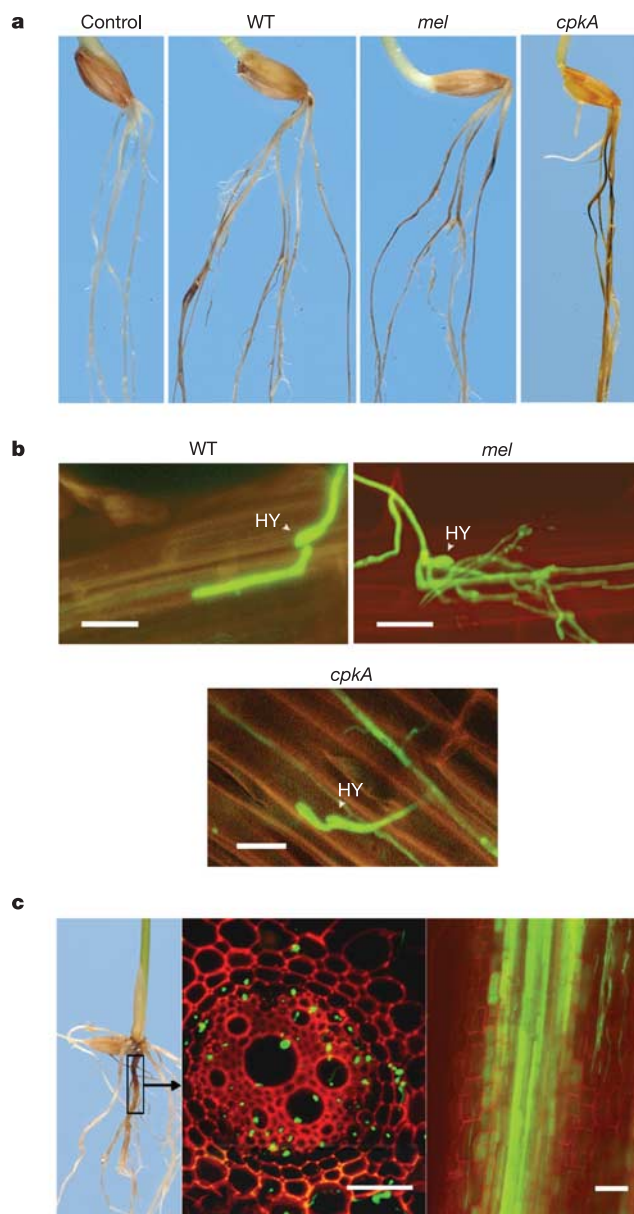


Figure 2 Infection of roots by wild type and mutant *M. grisea* strains. **a**, Roots of barley seedlings (cultivar Golden Promise) that have been mock-inoculated or infected with the *M. grisea* wild type (WT) or mutant (*mel*, *cpkA*) strains. **b**, Formation of hyphopodia-like structures (HY) and invasive growth within epidermal cells during the early stages of infection. **c**, *M. grisea* penetrates the stele. Confocal imaging of radial and longitudinal sections of a three-week-old rice seedling (cultivar Nipponbare) infected with GFP-tagged *M. grisea* (strain Guy11). Scale bars, 25 μ m (**b**), 40 μ m (**c**).

of differentiation events that occurs on leaves and that are typical of root pathogens. The further genetic dissection of these processes and their regulation represents an exciting future challenge. This infection system also provides an opportunity to analyse the role of

genes in the infection process independently of the formation of appressoria. From a practical perspective, our experiments indicate that soil-borne inoculum and root infection may be important for *M. grisea* epidemiology. We have provided clear evidence to validate the biological relevance of the root infection process for plant infection under laboratory conditions. In the field, the major means of dissemination of *M. grisea* is likely to be through aerial dispersal of conidia, although seeds and crop residue have also been implicated as sources of inoculum²². Our data suggest that root infection may also contribute to primary infection and disease establishment in the field. The epidemiological significance of root infection for rice blast disease remains to be addressed. However this possibility merits serious attention, particularly as *M. grisea* is closely related to major root-infecting pathogens of substantial economic significance such as the take-all fungus. Regardless of the current significance of the root infection process in the field, the threat of changes in behaviour of the pathogen as a consequence of altered cultural practices associated with intensification of rice production is also an issue of major concern. Investigation of the evolutionary events that have led to the ability of different members of the *Magnaporthaceae* to infect roots and aerial tissues is likely to be a fruitful avenue for further research. Interestingly, two other

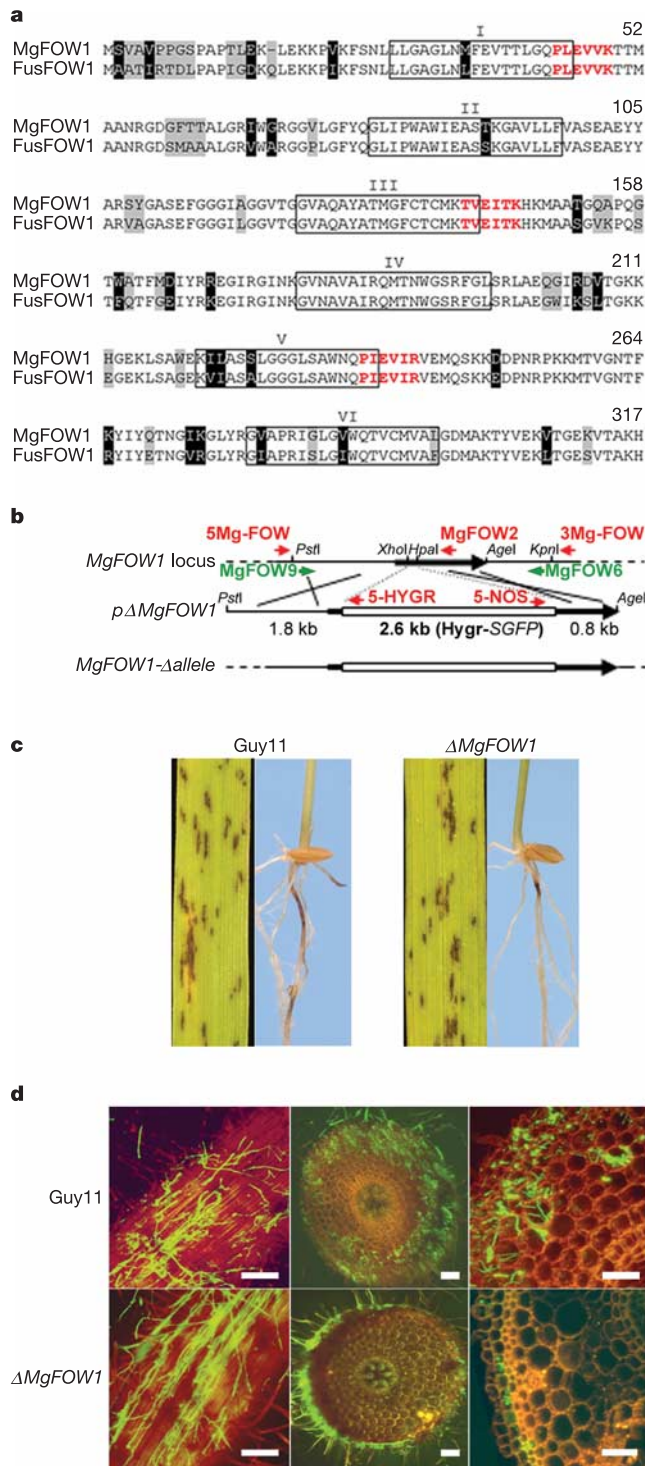


Figure 3 Colonization of rice roots by *M. grisea* is dependent on *MgFOW1*. **a**, Alignment of *MgFOW1* and *F. oxysporum* *FOW1*. Amino acids that differ are boxed (black, conserved; grey, non-conserved); conserved motifs of mitochondrial carrier proteins are indicated in red. Segments I–VI represent putative transmembrane regions. **b**, *MgFOW1* gene replacement. Primers for verification of gene replacement and complementation are indicated in red and green, respectively. **c**, Infection of rice (cultivar Nipponbare) by the Guy11 wild type and a Δ *MgFOW1* mutant. **d**, Confocal images of two-week-old rice seedlings (cultivar Nipponbare) infected with Guy11 and Guy11- Δ *MgFOW1* strains. Scale bar, 40 μ m.

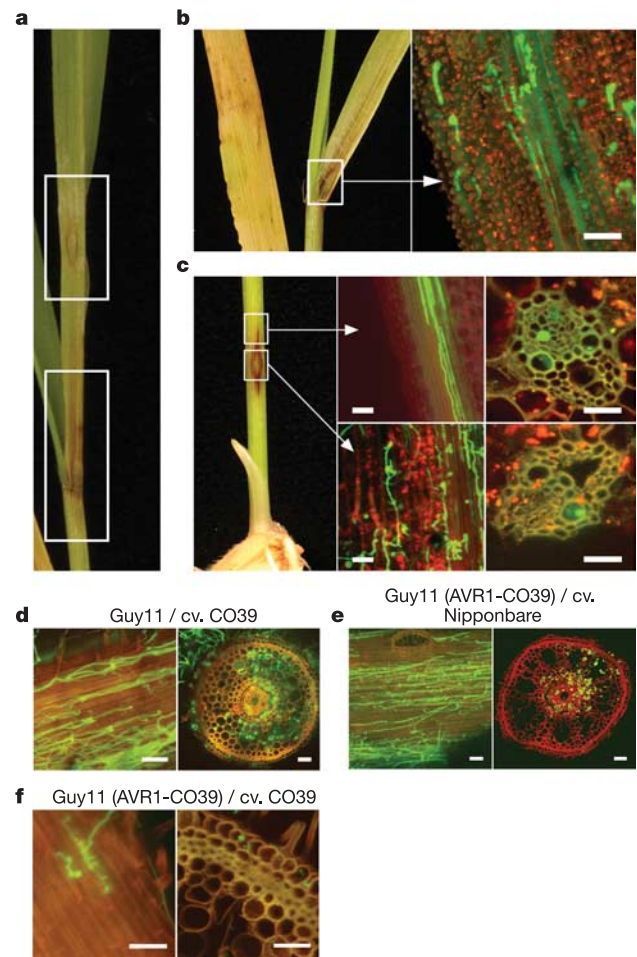


Figure 4 *M. grisea* can spread systemically from the roots to the aerial tissues and cause typical blast symptoms. **a–c**, Four-week-old root-infected rice seedlings (cultivar Nipponbare) showing disease symptoms on the leaf (upper box) and collar (lower box) (**a**). Disease symptoms on the collar (**b**) and stem (**c**) with confocal images showing GFP-expressing *M. grisea* Guy11 in the diseased areas and also in the vascular tissue of the leaf and stem. **d–f**, *Pi-CO39(t)*-mediated specific disease resistance operates in rice roots. Confocal microscopy of compatible (**d**, **e**) and incompatible (**f**) interactions. Cultivar, cv. Scale bar, 40 μ m.

pathogens that are traditionally regarded as being restricted to the aerial plant tissues have recently been reported to infect the roots of their hosts^{23,24} (*Leptosphaeria maculans*, which causes stem canker of brassicas and *Cercospora beticola*, which causes leaf spot disease of sugar beet). Taken together these observations suggest that soil-borne inoculum and root infection may be significant components of the life cycles of diseases that are traditionally regarded as only affecting above-ground parts of plants. This has important implications for the development of new strategies for plant breeding and disease control. □

Methods

DNA manipulations, fungal growth conditions and plant infection experiments

Standard molecular biology procedures were followed for cloning and enzymatic manipulations with DNA²⁵. Amino acid sequence comparisons and alignments were carried out using vector NTI Suite8 (InforMax, Invitrogen). Fungal strains were cultured and maintained as described¹⁴. The plant material used was rice (*Oryza sativa* L.) cultivars Nipponbare (ssp. *japonica*) and CO39 (ssp. *indica*), barley (*Hordeum vulgare*) cultivar Golden Promise and wheat (*Triticum aestivum*) cultivar Riband. Leaf and root infection assays were carried out as described^{10,14}.

Construction of SGFP binary vectors and *Agrobacterium*-mediated transformation

A modified pCambia1300 (Cambia, Canberra, Australia) binary vector containing a hygromycin resistance gene and the SGFP gene (encoding a GFP variant that contains a serine-to-threonine substitution at amino acid 65) from pCT74 (ref. 26) was constructed. The resulting vector (pCambiaGfp) was introduced into *Agrobacterium tumefaciens* strain AGL-1 and transformed into *M. grisea* as described²⁷. Transformants expressing SGFP were selected under ultraviolet light. A second binary vector expressing SGFP and containing the *ILV1* gene (which confers resistance to sulfonylurea) (pSULF) was also constructed by ligating *XhoI*/EcoRI-digested pCB1532 (ref. 28) into pCambia1300. The SGFP gene from pCT74 was then ligated into *Sall*/BamHI-digested pSULF to give pSULFGfp.

Microscopy

Infected root samples were stained with chlorazole black E as described²⁹. Confocal optical section stacks of infected plant material were collected using a Leica TCS-NT confocal microscope. SGFP fluorescence was detected with a 515 nm bandpass emission filter and autofluorescence of the plant cell walls with a 595 nm bandpass emission filter.

Construction of *MgFOW1* gene replacement vector

The candidate *FOW1* homologue *MgFOW1* (annotated at locus n. MG07201.4, chromosome I, contig 2.1337 in the *M. grisea* genome database, <http://www.broad.mit.edu/annotation/fungi/magnaporthe/>) was amplified with primers 5Mg-FOW (5'-CGGTGTCTCTCTGAGTACG-3') and 3Mg-FOW (5'-CTATGCC TCTGGTTCTATACCG-3') from *M. grisea* Guy11 strain. The sequence of the 3.7 kilobase (kb) amplification product was identical to that in the genome database (from *M. grisea* 70-15 strain). The polymerase chain reaction (PCR) product was digested with *PstI* and *KpnI* and the resulting 3.4 kb restriction fragment cloned into pBluescriptSK⁺ (Stratagene). A 170 bp *XhoI*-*HpaI* fragment within the predicted ORF was replaced by the ~2.6 kb *XhoI*-*SmaI* fragment from pCambiaGfp containing the hygromycin resistance cassette and the SGFP gene. An ~5.2 kb *PstI*-*KpnI* fragment encompassing the 2.6 kb *XhoI*-*SmaI* fragment and flanking DNA sequences was digested with *AgeI* and subcloned into pCambia1300. This gene replacement construct was then introduced into the *M. grisea* strain Guy11 by *Agrobacterium*-mediated transformation. The transferred DNA insertion sites were confirmed by sequencing the PCR fragments amplified using the primer pairs: 5Mg-FOW and MgFOW2 (5'-CTTCTTGCCGGTGACATCGCG-3'), 5Mg-FOW and 5-HYGR (5'-GCCGATAGTGGAACCCGACGC-3'), 5-NOS (5'-CTAGA TCCGATGATAAGCTGTC-3') and 3Mg-FOW (see above).

For complementation experiments the 3.4 kb *PstI*-*KpnI* fragment containing the intact *MgFOW1* gene was blunt-ended, cloned into *SmaI*-digested pSULF and introduced into the Δ *MgFOW1* mutants by *Agrobacterium*-mediated transformation. Transformants were selected with sulfonylurea (Greyhound, UK) and analysed by PCR with the primers MgFOW9 5'-GCAGTACTGTGATGACTAGG-3' and MgFOW6 (5'-GGTTTCGAGTGG CCGCTGTCG-3'), which yielded a ~2.8 kb amplification product fragment derived from the *MgFOW1* locus.

Received 28 April; accepted 20 July 2004; doi:10.1038/nature02880.

- Mendgen, K., Hahn, M. & Deising, H. Morphogenesis and mechanisms of penetration by plant pathogenic fungi. *Annu. Rev. Phytopathol.* **34**, 367–386 (1996).
- Tucker, S. L. & Talbot, N. J. Surface attachment and pre-penetration stage development by plant pathogenic fungi. *Annu. Rev. Phytopathol.* **39**, 385–418 (2001).
- Talbot, N. J. On the trail of a cereal killer: exploring the biology of *Magnaporthe grisea*. *Annu. Rev. Microbiol.* **57**, 177–202 (2003).
- International Rice Blast Genome Consortium. RICEBLAST.ORG, <http://www.riceblast.org/> (2003).
- Cannon, P. F. The newly recognized family Magnaporthaceae and its interrelationships. *Systema Ascomycetum* **13**, 25–42 (1994).
- Hornby, D. (ed.) *Biology and Control of Take-all* (Academic Press Inc., London, 1981).
- Hornby, D. (ed.) *Take-all Disease of Cereals: a Regional Perspective* (CAB International, Wallingford, 1998).

- Landschoot, P. J. & Jackson, N. *Magnaporthe poae* sp. nov., a hyphopodiate fungus with a *Phialophora* anamorph from grass roots in the United States. *Mycol. Res.* **93**, 59–62 (1989).
- Scott, D. B. & Deacon, J. W. *Magnaporthe rhizophila* sp. nov., a dark mycelial fungus with a *Phialophora* conidial state, from cereal roots in South-Africa. *Trans. British Mycol. Soc.* **81**, 77–81 (1983).
- Dufresne, M. & Osbourn, A. E. Definition of tissue-specific and general requirements for plant infection in a phytopathogenic fungus. *Mol. Plant Microbe Interact.* **14**, 300–307 (2001).
- De Jong, J. C., McCormack, B. J., Smirnov, N. & Talbot, N. J. Glycerol generates turgor in rice blast. *Nature* **389**, 244–245 (1997).
- Howard, R. J. & Valent, B. Breaking and entering: host penetration by the fungal rice blast pathogen *Magnaporthe grisea*. *Annu. Rev. Microbiol.* **50**, 491–512 (1996).
- Rodrigues, F. A., Benhamou, N., Datnoff, L. E., Jones, J. B. & Belanger, R. R. Ultrastructural and cytochemical aspects of silicon-mediated rice blast resistance. *Phytopathology* **93**, 535–546 (2003).
- Talbot, N. J., Ebbole, D. J. & Hamer, J. E. Identification and characterization of *MPG1*, a gene involved in pathogenicity from the rice blast fungus *Magnaporthe grisea*. *Plant Cell* **5**, 1575–1590 (1993).
- Chumley, F. G. & Valent, B. Genetic analysis of melanin deficient, nonpathogenic mutants of *Magnaporthe grisea*. *Mol. Plant Microbe Interact.* **3**, 135–143 (1990).
- Valent, B., Farrall, L. & Chumley, F. *Magnaporthe grisea* genes for pathogenicity and virulence identified through a series of backcrosses. *Genetics* **127**, 87–101 (1991).
- Mitchell, T. K. & Dean, R. A. The cAMP-dependent protein kinase catalytic subunit is required for appressorium formation and pathogenesis by the rice blast pathogen *Magnaporthe grisea*. *Plant Cell* **7**, 1869–1878 (1995).
- Xu, J. R., Urban, M., Sweigard, J. A. & Hamer, J. E. The CPKA gene of *Magnaporthe grisea* is essential for appressorial penetration. *Mol. Plant Microbe Interact.* **10**, 187–194 (1997).
- Inoue, I., Namiki, F. & Tsuge, T. Plant colonization by the vascular wilt fungus *Fusarium oxysporum* requires *FOW1*, a gene encoding a mitochondrial protein. *Plant Cell* **14**, 1869–1883 (2002).
- Farman, M. L. et al. Analysis of the Structure of the AVR1-CO39 avirulence locus in virulent rice-infecting isolates of *Magnaporthe grisea*. *Mol. Plant Microbe Interact.* **15**, 6–16 (2002).
- Tian, D., Traw, M. B., Chen, J. Q., Kreitman, M. & Bergelson, J. Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. *Nature* **423**, 74–76 (2003).
- Ou, S. H. *Rice Diseases* (Commonwealth Mycological Institute, Kew, Surrey, UK, 1985).
- Sosnowski, M., Ramsey, M., Murray, G., Scott, E. & Wilmshurst, C. Symptoms of blackleg (*Leptosphaeria maculans*) on the roots of canola in Australia. *Plant Pathol.* **50**, 808 (2001).
- Vereijssen, J., Schneider, H. J. M. & Termorshuizen, A. J. Possible root infection of *Cercospora beticola* in sugar beet. *Eur. J. Plant Pathol.* **110**, 103–106 (2004).
- Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: a Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989).
- Lorang, J. M. et al. Green fluorescent protein is lighting up fungal biology. *Appl. Environ. Microbiol.* **67**, 1987–1994 (2001).
- Rho, H. S., Kang, S. & Lee, Y. H. *Agrobacterium tumefaciens*-mediated transformation of the plant pathogenic fungus, *Magnaporthe grisea*. *Mol. Cells* **12**, 407–411 (2001).
- Sweigard, J. A., Chumley, F., Carroll, A., Farrall, L. & Valent, B. A series of vectors for fungal transformation. *Fungal Genet. Newsl.* **44**, 52–53 (1997).
- Resendes, C. M., Geil, R. D. & Guinel, F. C. Mycorrhizal development in a low nodulating pea mutant. *New Phytol.* **150**, 563–572 (2001).
- Gangopadhyay, S. & Row, K. V. K. Perennation of *Pyricularia oryzae* briosi et cav. in sclerotial state. *Int. J. Trop. Plant Dis.* **4**, 187–192 (1986).

Acknowledgements We thank the *M. grisea* community for providing fungal strains; L. Ciuffetti for the pCT74 reporter vector, and J. Deacon and R. Gutteridge for providing the images for Fig. 11 and 1m, respectively. A.S. has been supported by a Marie Curie fellowship from the European Community and by the Gatsby Charitable Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.E.O. (annie.osbourn@sainsbury-laboratory.ac.uk).

Structural basis for packaging the dimeric genome of Moloney murine leukaemia virus

Victoria D'Souza & Michael F. Summers

Howard Hughes Medical Institute and Department of Chemistry and Biochemistry, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, Maryland 21250, USA

All retroviruses specifically package two copies of their genomes during virus assembly, a requirement for strand-transfer-mediated recombination during reverse transcription^{1,2}. Genomic RNA exists in virions as dimers, and the overlap of RNA elements that promote dimerization and encapsidation suggests

that these processes may be coupled^{3–5}. Both processes are mediated by the nucleocapsid domain (NC) of the retroviral Gag polyprotein³. Here we show that dimerization-induced register shifts in base pairing within the Ψ -RNA packaging signal of Moloney murine leukaemia virus (MoMuLV) expose conserved UCUG elements that bind NC with high affinity (dissociation constant = 75 ± 12 nM). These elements are base-paired and do not bind NC in the monomeric RNA. The structure of the NC complex with a 101-nucleotide 'core encapsidation' segment of the MoMuLV Ψ site⁶ reveals a network of interactions that promote sequence- and structure-specific binding by NC's single CCHC zinc knuckle. Our findings support a structural RNA switch mechanism for genome encapsidation, in which

protein binding sites are sequestered by base pairing in the monomeric RNA and become exposed upon dimerization to promote packaging of a diploid genome.

The MoMuLV is a prototypical retrovirus widely used in human gene therapy trials and extensively studied as a model for retrovirus assembly and genome encapsidation⁷. The 350-nucleotide MoMuLV packaging signal (Ψ , Fig. 1a) forms a structurally organized domain that independently directs dimerization and packaging of the viral RNA⁸. The secondary structure of the Ψ site changes on dimerization⁹, and it has been suggested that RNA conformational changes may help regulate genome packaging and other genome-related replication events^{9,10}. Dimerization is promoted by two stem loops that form intermolecular duplexes (DIS-1, A204–G229; and

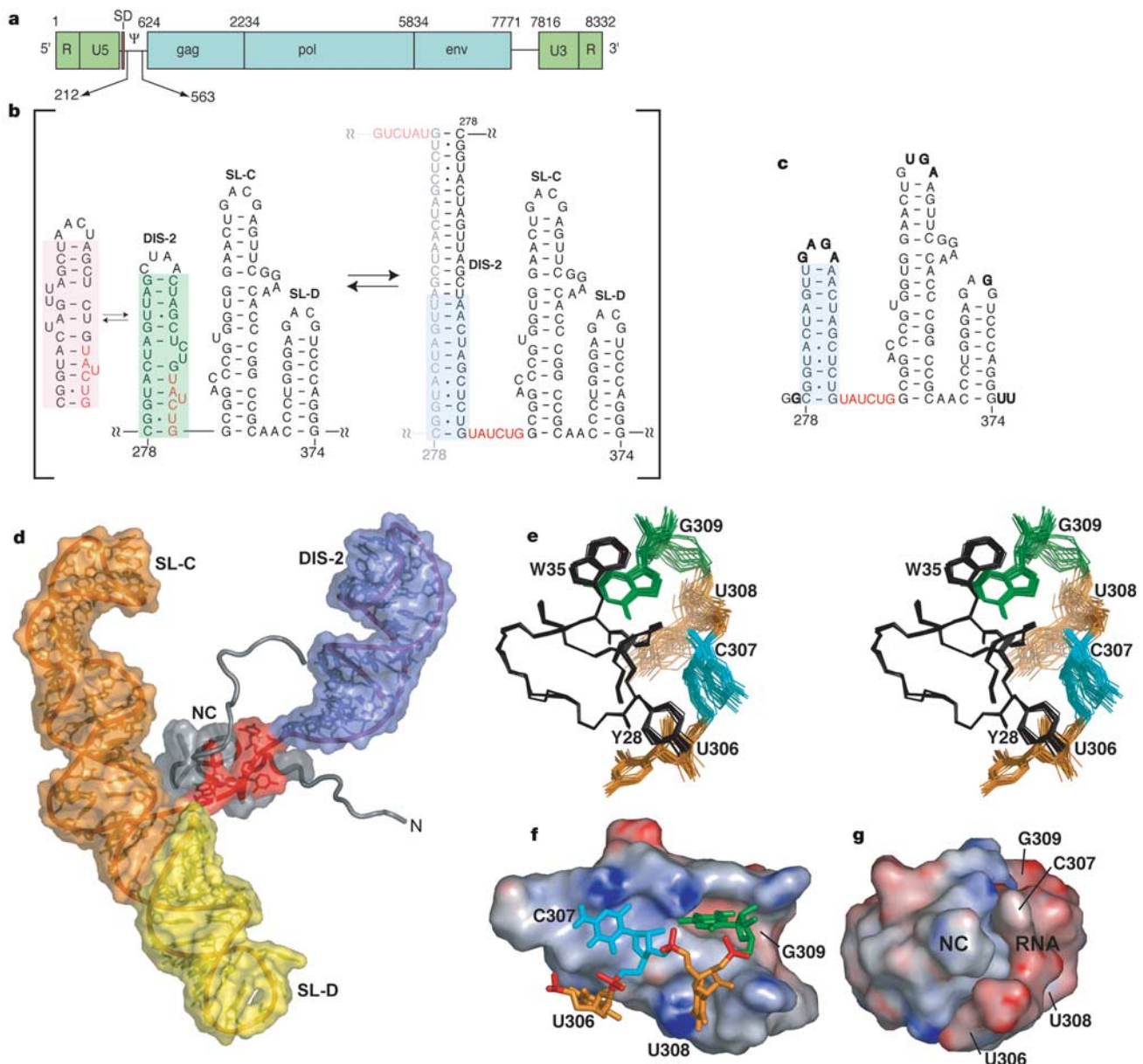


Figure 1 MoMuLV Ψ site and structure of the NC-m Ψ^{CES} complex. **a**, Unspliced MoMuLV genome showing the location of the Ψ packaging signal. **b**, Secondary structure of the core encapsidation signal (Ψ^{CES}). DIS-2 exists in two alternate monomeric conformations (shaded red and green), and undergoes a frame shift upon dimerization (shaded blue) that exposes a UAUCUG element (red). **c**, Secondary structure of m Ψ^{CES} with non-native nucleotides shown in bold. DIS-2 base pairings match those of the dimeric form of native Ψ^{CES} . **d**, Representative NC-m Ψ^{CES} structure DIS-2 (blue), SL-C (orange) and SL-D

(yellow), the UCUG segment (red) and NC (grey). **e**, Stereo image showing the best-fit backbone superposition of the CCHC zinc knuckle for the 20 calculated NC-m Ψ^{CES} structures. **f**, Interactions between the zinc knuckle (coloured according to electrostatic surface potential) and the U306CUG element: U (orange), C (cyan) and G (green). **g**, Surface representations of the zinc knuckle and U306CUG segment showing complementarity of shape and electrostatic potential at the protein–RNA interface.

DIS-2, C278–G309)^{9,11,12} and two additional stem loops that can form ‘kissing complexes’ through intermolecular base pairing of their tetraloop nucleotides (SL-C, G310–C352; and SL-D, C355–G374)¹³. Although all four stem loops are required for efficient encapsidation, fragments of Ψ containing SL-C, SL-D and portions of DIS-2 are capable of independently directing the packaging of heterologous RNAs into virus-like particles⁶. These residues comprise a ‘core encapsidation signal’ (Ψ^{CES} ; Fig. 1b; ref. 6) that can bind NC stoichiometrically and with high affinity¹⁴. The DIS-2 stem loop of Ψ^{CES} undergoes a register-shift in base pairing upon dimerization⁹, in which residues U304–G309 are base paired in the monomer but not in the dimer (Fig. 1b). To gain insights into the molecular determinants of genome packaging, we determined the structure of the complex between NC and a 101 nucleotide Ψ^{CES} mutant, engineered to remain monomeric in solution but retain internal base pairing of the dimer^{14,15} ($m\Psi^{\text{CES}}$; Fig. 1c).

The NC– $m\Psi^{\text{CES}}$ structure demonstrated an unexpected binding mode, in which the single CCHC zinc knuckle of NC interacts with the UCUG linker that connects DIS-2 to SL-C (Fig. 1d–f). The binding interface exhibits significant complementarity of shape and charge (Fig. 1g). U306 and C307 pack against the side chain of Tyr 28, and U308 packs against the side chains of Ala 27 and Leu 21. The binding of these three nucleotides appears to be promoted by the following direct or water mediated interactions: U306–5′-phosphodiester to Tyr 28–OH, U306–O4 to Lys 42–NH₃⁺, C307–N3 and/or –O2 atoms to Lys 42–NH₃⁺, U308–O2′ to Arg 18 guanidinium group, U308–O4 to Lys 30–NH₃⁺ and U308–5′-phosphodiester to Lys 37–NH₃⁺. The G309 nucleotide base fits deeply into a pocket defined by the side chains of Leu 21, Ala 27, Trp 35 and Ala 36, and forms hydrogen bonds with backbone NH and O atoms located at the bottom of the pocket (G309–O6–Ala 27–NH; G309–O6–Ala 36–NH; G309–N1H–Gln 25–O; G309–NH21–Gln 25–O; Fig. 2a). Similar interactions with exposed guanines have been observed in human immunodeficiency virus Type-1 (HIV-1) NC–RNA^{16,17} and zinc knuckle–DNA¹⁸ complexes (Fig. 2b), although for HIV-1 NC, tight binding requires two adjacent CCHC zinc knuckles, each of which interacts directly with only a single nucleotide base (G)^{16,17}. Also, the HIV-1 pockets are ‘pre-formed’, being derived from residues on the surface of the folded zinc knuckle domain. In contrast, the guanosine binding pocket of the MoMuLV zinc knuckle is only partially formed in the absence of $m\Psi^{\text{CES}}$, with residues Arg 18–Asp 24 being disordered in the free protein and folding only on binding to the RNA. The remaining residues of the amino-terminal tail (Ala 1–Arg 17), and the carboxy-terminal tail (Arg 44–Leu 56), remain disordered on binding and do not interact tightly or specifically with the RNA. These findings are consistent with mutagenesis studies showing that basic residues between Arg 16 and Pro 43 are required for efficient genome packaging, whereas those in the flexible regions of NC (including the entire flexible portion of the C-terminal tail) are disposable^{4,19}.

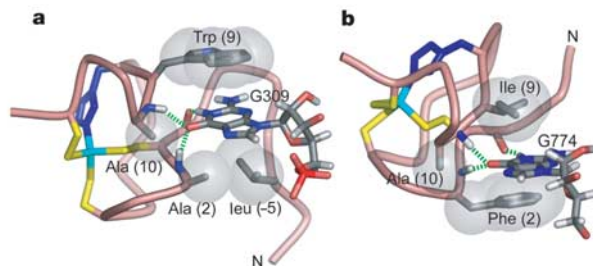


Figure 2 Comparison of the guanosine binding sites of the MoMuLV and HIV-1 zinc knuckles. **a**, MoMuLV zinc knuckle in NC– $m\Psi^{\text{CES}}$. **b**, HIV-1 zinc knuckle in NC–SL3 (ref. 16). For comparison, the numbering scheme used for both zinc knuckles begins with the first cysteine labelled Cys 1 (this figure only). The zinc atom (cyan) and cysteine (yellow) and histidine (blue) side chains are shown.

NMR chemical shifts and nuclear Overhauser effect (NOE) cross peak patterns of the DIS-2, SL-C and SL-D stem loops were not perturbed by NC binding, and their structures are identical to those observed for the free $m\Psi^{\text{CES}}$ RNA¹⁵. To determine if transient interactions with these elements might contribute to binding, NC titration experiments were performed with truncated forms of $m\Psi^{\text{CES}}$. Tight binding (dissociation constant (K_d) < 200 nM) was observed for all fragments that included the linker residues, including the shortest fragment tested, r-UAUCUG (K_d = 75 ± 12 nM; Supplementary Fig. S1). For comparison, NC binds native Ψ^{CES} with a dissociation constant of 132 ± 55 nM (ref. 14). Binding was not observed for any RNAs that lacked the linker¹⁴.

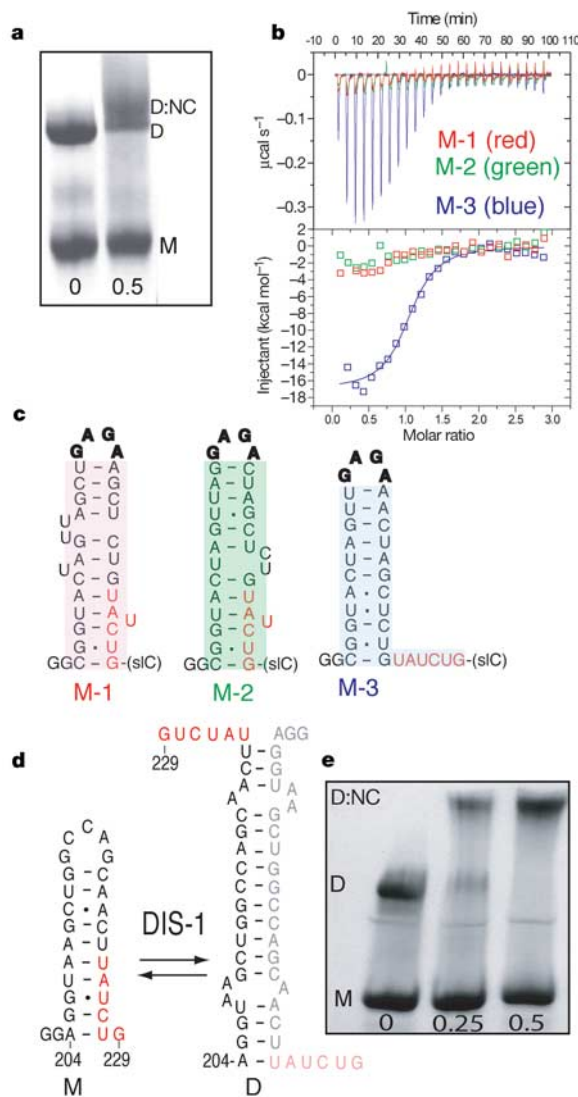


Figure 3 NC binds to dimeric forms of DIS-1 and DIS-2. **a**, Native PAGE data obtained for an RNA construct corresponding to residues G276–C352 of native Ψ^{CES} (DIS2–C) in the absence (left) and presence (right) of NC (0.5 equivalents). The titration was performed under equilibrium conditions of monomeric (M) and dimeric (D) species, and a band shift was observed only for the dimeric species. **b**, **c**, ITC NC titration results obtained for DIS2–C constructs containing GNRA-type mutations (bold) that stabilize base pairings of the monomeric (M-1 and M-2) and dimeric (M-3) Ψ site. All three constructs are monomeric under experimental conditions, based on native gel electrophoresis. Base pairings were confirmed by 2D ¹H NMR. Colour shadings and secondary structures correspond to those shown in Fig. 1b for the native Ψ site. NC binds M-3 with significant affinity (K_d = 173 ± 32 nM), but does not interact tightly with M-1 or M-2. **d**, Secondary structures determined by 2D ¹H NMR for a 28 nucleotide DIS-1 RNA in its monomeric (M) and dimeric (D) forms. **e**, Native PAGE data showing that only the dimeric form of DIS-1 binds NC.

In addition, NMR chemical shifts and NOE cross peak patterns in two-dimensional (2D) NOESY data obtained for the NC-r(UAUCUG) complex matched the data obtained for NC-m Ψ^{CES} complex (Supplementary Figs S2 and S3). These data indicate that the UCUG linker element, which is conserved among the murine C-type retroviruses¹⁵, is both necessary and sufficient for high affinity NC binding. DIS-2 contains a second UCUG sequence that is also conserved (U300CUG), but this element is sequestered by intramolecular base pairing and stacking interactions and does not interact with NC.

UCUG sequences are unusually abundant in the region between the primer binding site (PBS) and Gag initiation codon of the mammalian C-type retroviruses, occurring at a frequency of approximately 1 in 50 nucleotides. By comparison, these sequences occur at a frequency of 1 in 225 nucleotides in the coding and LTR (long terminal repeat) regions of the MoMuLV and other retroviral genomes reported in the NIH NCBI Gene Bank. Thirteen UCUG sequences are present in the 418-nucleotide segment between the PBS and Gag start sites of the MoMuLV genome, accounting for more than 12% of the nucleotides in this region. Five of these sequences are located within the 55 nucleotides between Ψ and the Gag initiation codon, and this may explain why ' Ψ +' RNAs (that is, RNAs with a 3'-extended Ψ site) package heterologous RNAs with greater efficiency than Ψ -only RNAs²⁰. Not all of the UCUG segments are likely to bind NC owing to their participation in intramolecular interactions (as observed for U300CUG). However, the high abundance of these sequences is consistent with proposals that the NC domains of assembling Gag molecules bind cooperatively to multiple sites within the Ψ -RNA packaging signal. No UCUG sequences were found between the PBS and Gag initiation codons of representative lentivirus (HIV-1) and D-type retrovirus (Mason-Pfizer monkey virus) genomes, and fewer than three UCUG sequences were observed in representative B-type (mouse mammary tumour virus), avian C-type (Rous sarcoma virus) and HTLV/BLV (human T-lymphotrophic virus-I) retrovirus genomes. The NC proteins of these retroviruses contain two CCHC zinc knuckle domains with amino acid compositions that differ significantly from those of the mammalian C-type retroviruses. These differences may explain why chimeric HIV-1 virions containing the MoMuLV NC domain are able to efficiently package the MoMuLV genome but unable to package the HIV-1 genome²¹, and are consistent with the observation that retroviruses of a given genus are often capable of packaging each other's genomes²².

Previous chemical accessibility mapping and free energy calculations suggested that base pairing within the DIS-2 stem loop of the intact MoMuLV Ψ site changes upon dimerization⁹. NMR data

obtained for RNAs containing the native DIS-2 sequence confirmed the predicted structures shown in Fig. 1b (Supplementary Fig. S4). The fact that residues U306–G309 are internally base-paired in the monomeric hairpin and exposed in the duplex⁹ raised the possibility that NC might bind preferentially to the duplex. To test this hypothesis, a 77 nucleotide RNA corresponding to stem loops DIS-2 and SL-C (DIS2-C) of the native Ψ^{CES} RNA was prepared for NC-binding studies. Titration of DIS2-C with NC resulted in a small but distinguishable electrophoretic band shift for the dimeric species, but did not affect the mobility of the monomer (Fig. 3a; ref. 15). To quantitatively test for structure-dependent NC binding, isothermal titration calorimetry (ITC) data were obtained for mutants of DIS2-C containing GNRA sequences that specifically stabilize base pairings observed in the monomeric and dimeric Ψ site (see Figs 1b and 3b, c). As shown in Fig. 3b, RNAs with base pairings of the monomer (M1 and M2 in Fig. 3c) did not bind NC, whereas the construct with base pairings of the dimer (M3) bound NC with high affinity. These findings confirm that NC binds only to the dimeric form of DIS-2, in which the U306CUG element is exposed.

The MoMuLV Ψ site contains a second palindromic segment that promotes genome dimerization *in vitro*^{11,12} and encapsidation *in vivo*¹², and this segment (DIS-1, A204–G229) has also been predicted to undergo a register shift in base pairing on conversion from a monomeric hairpin to a dimeric duplex¹¹. DIS-1 also contains a 3'-UCUG sequence that could potentially bind NC. A 28-nucleotide RNA with sequence of DIS-1 (G202–G229, Fig. 3d) forms monomers and dimers at RNA concentrations above 0.1 mM, and titration of DIS-1 with NC resulted in a significant electrophoretic band shift for the dimeric (but not the monomeric) species (Fig. 3e). 2D NOESY experiments confirmed that the UCUG segment is base-paired in the hairpin and exposed in the duplex, and that NC binding is mediated by interactions between the CCHC zinc knuckle and the exposed UCUG element of the duplex (data not shown). Thus, DIS-1 exhibits dimerization-dependent structural and NC-binding behaviour very similar to that observed for DIS-2.

Our findings support a structural RNA switch mechanism for genome encapsidation, in which conserved UCUG elements associated with DIS-1 and DIS-2 are sequestered in the monomeric RNA and become exposed to promote high affinity NC binding on dimerization (Fig. 4). In both cases, asymmetry within the pseudo-palindromic stem loops allows dimerization-induced register shifts that re-optimize base pairing in the dimer and expose the protein binding site. These findings are consistent with studies suggesting that genome dimerization and encapsidation events are intimately coupled^{5,7,23–27}. Conserved UCUG elements downstream of SL-D that exhibit dimerization-dependent chemical reactivity (U510CUG and U559CUG)⁹ may also bind NC in a dimerization-dependent manner. Further studies of these and the five UCUG elements in the segment downstream of Ψ that is required for optimal packaging efficiency²⁰ are warranted. Dimerization-dependent conformational changes have recently been detected in the HIV-1 Ψ site¹⁰, and it seems that at least one high affinity NC binding site that is present in the dimeric Ψ site (stem loop SL2)¹⁷ exists in a significantly different conformation in the monomeric RNA. Conformational RNA switch mechanisms may therefore be commonly used by retroviruses to promote specific packaging of a diploid genome. □

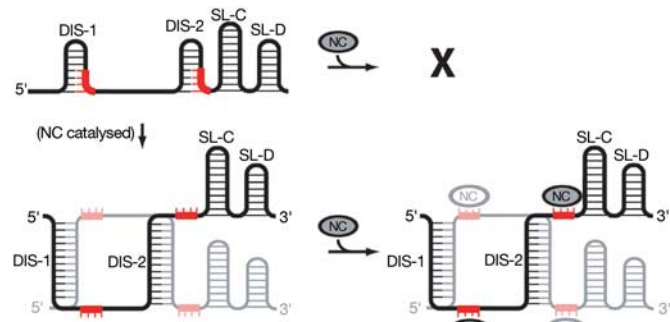


Figure 4 Structural changes in the Ψ -site that seem to serve as a switch for the selective binding of NC to the dimer. Conserved UCUG bases (red) are base paired in the monomeric state and become exposed for NC binding upon dimerization. Dimerization may promote exposure of additional downstream UCUG elements to enhance the specific packaging of a diploid genome.

Methods

RNA design

The GAGC tetraloops of SL-C and SL-D were mutated to GUGA and GAGG, respectively, to eliminate intermolecular kissing interactions that can lead to aggregation at concentrations required for NMR studies¹⁴. These mutations do not affect the structures of the stem loops or the NC binding properties of the RNA^{14,15}. In addition, the palindromic AGCU segment of DIS-2 was replaced by GAGA, a GNRA-type tetraloop that stabilizes hairpin structures²⁸. This mutation stabilizes base pairings in the stem of DIS-2 that match those observed for the native, dimeric Ψ site (see Fig. 1b, c). The resulting m Ψ^{CES} RNA remains monomeric at concentrations well above 1.0 mM (based on native

gel electrophoresis and NMR) and binds NC with the same affinity and stoichiometry observed for the native Ψ^{CES} RNA¹⁴.

Sample preparation

MoMuLV NC protein and RNA constructs were prepared as described^{14,15}. RNAs of 35 nucleotides or less were obtained from Dharmacon and purified by denaturing gel electrophoresis. Samples for all NMR, ITC and polyacrylamide gel electrophoresis (PAGE) measurements were prepared in Tris-HCl buffer (10 mM at pH 7.0, 10 mM NaCl, 0.1 mM ZnCl₂ and 0.1 mM β -mercaptoethanol).

NC binding experiments

ITC data (VP-ITC calorimeter, MicroCal Corp.) were measured at 30 °C. Exothermic heats of reaction were measured for 25 injections of NC protein (80 μ M) into 1.4 ml of RNA (5.0 μ M). Binding curves were analysed by nonlinear least squares fitting of the baseline corrected data to a single binding site model as described¹⁴. RNAs for PAGE experiments (0.5 mM) were heated for 2 min at 90 °C and cooled on ice before addition of NC. Samples were loaded onto native polyacrylamide gels and electrophoresed at 4 °C in Tris-borate buffer (45 mM Tris base, 45 mM boric acid, pH 8.3), and gels were stained with Stains-All (Sigma).

NMR spectroscopy and signal assignments

NMR data (Bruker DRX spectrometer, 800 MHz ¹H, T = 15, 25 and 35 °C) were obtained from a combination of two-, three- and four-dimensional NOESY data, for samples with combinations of natural abundance and ¹⁵N-, ¹³C-labelled NC and both ¹⁵N-, ¹³C-labelled and perdeuterated/selectively protonated RNA. RNA signals were assigned as described for the free m Ψ^{CES} RNA¹⁵. Protein backbone signals were assigned using standard triple resonance methods, and side chain signals were assigned from 3D and 4D ¹⁵N-, ¹³C-, and ¹⁵N/¹³C-edited NOESY data (see refs 16, 29, 30, and citations there-in). Titration of m Ψ^{CES} with NC resulted in NMR chemical shift changes and intermolecular ¹H-¹H NOEs for nucleotides U306–G309. All other m Ψ^{CES} NMR signals were unaffected by NC binding. Intermolecular NOEs were readily assigned from 2D NOESY data obtained for samples containing nucleotide-specifically protonated RNA (G^H-m Ψ^{CES} , A^H-m Ψ^{CES} , U^H-m Ψ^{CES} , C^H-m Ψ^{CES} ; containing protonated G, A, U and C, respectively, with the remaining nucleotides perdeuterated), in combination with 3D ¹³C-edited NOESY data obtained for RNA complexes with ¹⁵N-, ¹³C-labelled NC. Some RNA NMR signals were significantly broadened on titration with NC, and in these cases, intermolecular NOEs were identified or confirmed on the basis of exchange-mediated intermolecular NOEs with the unbound RNA in samples containing a ~30% excess of RNA (see Supplementary Fig. S2). Subsequent to solving the structure of the NC-m Ψ^{CES} complex, 2D NOESY data were obtained for the NC complex with r-UAUCUG (Supplementary Fig. S3), which enabled confirmation of assignments made for NC-m Ψ^{CES} but did not lead to the identification of additional intermolecular NOEs. Backbone NH signals for residues of the N- and C-terminal tails (Ala 1–Arg 17 and Arg 44–Leu 56, respectively) were sensitive to exchange with H₂O protons and exhibited random coil chemical shifts (as also observed for the C α carbons).

Structure calculations

Upper interproton distance bounds of 2.7, 3.3 and 5.0 Å were employed for NOE cross peaks of strong, medium and weak intensity respectively, which were qualitatively assessed following intensity normalization of the different NOE data sets. Structures were calculated in torsion angle space with CYANA (http://www.las.jp/index_eg.html) starting from random initial angles (involving both protein and RNA), and stem loops DIS-2, SL-C and SL-D were refined independently using dipolar couplings obtained for the isolated stem loops. This approach is valid because the NMR chemical shifts and NOE cross peak patterns and intensities of the isolated stem loops and those in the NC-m Ψ^{CES} complex were indistinguishable¹⁵. Statistical information and superposition images are provided in Supplementary Table S1 and Fig. S5, respectively. Structure figures were generated with PyMOL (<http://pymol.sourceforge.net>).

Received 8 July; accepted 16 August 2004; doi:10.1038/nature02944.

- Varmus, H. E. Form and function of retroviral proviruses. *Science* **216**, 812–820 (1982).
- Hu, W. S. & Temin, H. M. Retroviral recombination and reverse transcription. *Science* **250**, 1227–1233 (1990).
- Prats, A.-C. *et al.* Cis elements and trans-acting factors involved in dimer formation of murine leukemia virus RNA. *J. Virol.* **64**, 774–783 (1990).
- Housset, V., De Rocquigny, H., Roques, B. P. & Darlix, J.-L. Basic amino acids flanking the zinc finger of Moloney murine leukemia virus nucleocapsid protein NCp10 are critical for virus infectivity. *J. Virol.* **67**, 2537–2545 (1993).
- Paillart, J.-C., Shehu-Xhilaga, M., Marquet, R. & Mak, J. Dimerization of retroviral RNA genomes: an inseparable pair. *Nature Rev. Microbiol.* **2**, 461–472 (2004).
- Mougel, M. & Barklis, E. A role for two hairpin structures as a core RNA encapsidation signal in murine leukemia virus virions. *J. Virol.* **71**, 8061–8065 (1997).
- Berkowitz, R., Fisher, J. & Goff, S. P. RNA packaging. *Curr. Top. Microbiol. Immunol.* **214**, 177–218 (1996).
- Mann, R., Mulligan, R. C. & Baltimore, D. Construction of a retrovirus packaging mutant and its use to produce helper-free defective retrovirus. *Cell* **33**, 153–159 (1983).
- Tounketti, N. *et al.* Effect of dimerization on the conformation of the encapsidation psi domain of Moloney murine leukemia virus RNA. *J. Mol. Biol.* **223**, 205–220 (1992).
- Huthoff, H. & Berkhout, B. Two alternating structures of the HIV-1 leader RNA. *RNA* **7**, 143–157 (2001).
- Oroudjev, E. M., Kang, P. C. E. & Kohlstaedt, L. A. An additional dimer linkage structure in Moloney murine leukemia virus RNA. *J. Mol. Biol.* **291**, 603–613 (1999).
- Ly, H. & Parslow, T. G. Bipartite signal for genomic RNA dimerization in the Moloney murine leukemia virus. *J. Virol.* **76**, 3135–3144 (2002).

- Kim, C.-H. & Tinoco, I. Jr. A retroviral RNA kissing complex containing only two G–C base pairs. *Proc. Natl Acad. Sci. USA* **97**, 9396–9401 (2000).
- D'Souza, V. *et al.* Identification of a high-affinity nucleocapsid protein binding site within the Moloney murine leukemia virus Ψ -RNA packaging signal. Implications for genome recognition. *J. Mol. Biol.* **314**, 217–232 (2001).
- D'Souza, V., Dey, A., Habib, D. & Summers, M. F. NMR structure of the 101 nucleotide core encapsidation signal of the Moloney murine leukemia virus. *J. Mol. Biol.* **337**, 427–442 (2004).
- De Guzman, R. N. *et al.* Structure of the HIV-1 nucleocapsid protein bound to the SL3 Ψ -RNA recognition element. *Science* **279**, 384–388 (1998).
- Amarasinghe, G. K. *et al.* NMR structure of the HIV-1 nucleocapsid protein bound to stem-loop SL2 of the Ψ -RNA packaging signal. *J. Mol. Biol.* **301**, 491–511 (2000).
- Schuller, W., Dong, C.-Z., Wecker, K. & Roques, B.-P. NMR structure of the complex between the zinc finger protein NCp10 of Moloney murine leukemia virus and the single-stranded pentanucleotide d(ACGCC): Comparison with HIV-NCp7 complexes. *Biochemistry* **38**, 12984–12994 (1999).
- Rein, A., Harvin, D. P., Mirro, J., Ernst, S. M. & Gorelick, R. J. Evidence that a central domain of nucleocapsid protein is required for RNA packaging in murine leukemia virus. *J. Virol.* **68**, 6124–6129 (1994).
- Bender, M. A., Palmer, T. D., Gelinas, R. E. & Miller, A. D. Evidence that the packaging signal of Moloney murine leukemia virus extends into the gag region. *J. Virol.* **61**, 1639–1646 (1987).
- Berkowitz, R. D., Ohagen, A., Hoglund, S. & Goff, S. P. Retroviral nucleocapsid domains mediate the specific recognition of genomic viral RNAs by chimeric Gag polyproteins during RNA packaging *in vivo*. *J. Virol.* **69**, 6445–6456 (1995).
- Yang, S. & Temin, H. M. A double hairpin structure is necessary for the efficient encapsidation of spleen necrosis virus retroviral RNA. *EMBO J.* **13**, 713–726 (1994).
- Hibbert, C. S., Mirro, J. & Rein, A. mRNA molecules containing MLV packaging signals are encapsidated as dimers. *J. Virol.* (in the press).
- Levin, J. G., Grimley, P. M., Ramseur, J. M. & Berezsky, I. K. Deficiency of 60 to 70S RNA in murine leukemia virus particles assembled in cells treated with actinomycin D. *J. Virol.* **14**, 152–161 (1974).
- Sakuragi, J.-I., Shioda, T. & Panganiban, A. T. Duplication of the primary encapsidation and dimer linkage region of Human immunodeficiency virus type 1 RNA results in the appearance of monomeric RNA in virions. *J. Virol.* **75**, 2557–2565 (2001).
- Torrent, C., Bordet, T. & Darlix, J.-L. Analytical study of rat retrotransposon VL30 RNA dimerization *in vitro* and packaging in murine leukemia virus. *J. Mol. Biol.* **240**, 434–444 (1994).
- Méric, C. & Goff, S. P. Characterization of Moloney murine leukemia virus mutants with single-amino-acid substitutions in the Cys-His box of the nucleocapsid protein. *J. Virol.* **63**, 1558–1568 (1989).
- Antao, V. P., Lai, S. Y. & Tinoco, I. Jr. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.* **19**, 5901–5905 (1991).
- Wüthrich, K. *NMR of Proteins and Nucleic Acids* (John Wiley, New York, 1986).
- Kay, L. E., Clore, G. M., Bax, A. & Gronenborn, A. M. Four-dimensional heteronuclear triple-resonance NMR spectroscopy of interleukin-1 β in solution. *Science* **249**, 411–414 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the HHMI staff and colleagues at UMBC for technical support. This research was supported by a grant from the NIH to M.F.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.F.S. (summers@hhmi.umbc.edu).

Trigger factor in complex with the ribosome forms a molecular cradle for nascent proteins

Lars Ferbitz¹, Timm Maier¹, Holger Patzelt², Bernd Bukau², Elke Deuerling² & Nenad Ban¹

¹*Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule Hönggerberg (ETH Zürich), HPK Gebäude, CH-8093 Zürich, Switzerland*

²*Zentrum für Molekulare Biologie (ZMBH), Universität Heidelberg, Im Neuenheimer Feld 282, D-69120 Heidelberg, Germany*

During protein biosynthesis, nascent polypeptide chains that emerge from the ribosomal exit tunnel encounter ribosome-associated chaperones, which assist their folding to the native state^{1,2}. Here we present a 2.7 Å crystal structure of *Escherichia coli* trigger factor, the best-characterized chaperone of this type, together with the structure of its ribosome-binding

domain in complex with the *Haloarcula marismortui* large ribosomal subunit. Trigger factor adopts a unique conformation resembling a crouching dragon with separated domains forming the amino-terminal ribosome-binding 'tail', the peptidyl-prolyl isomerase 'head', the carboxy-terminal 'arms' and connecting regions building up the 'back'. From its attachment point on the ribosome, trigger factor projects the extended domains over the exit of the ribosomal tunnel, creating a protected folding space where nascent polypeptides may be shielded from proteases and aggregation. This study sheds new light on our understanding of co-translational protein folding, and suggests an unexpected mechanism of action for ribosome-associated chaperones.

In all organisms, messenger RNA-directed protein synthesis is achieved by the ribosome. The atomic structures of the large (50S) and small (30S) ribosomal subunits and the structural study of the entire bacterial ribosome provided important insights into this sophisticated macromolecular machine³. The structure of the *H. marismortui* 50S subunit revealed a tunnel of approximately 100 Å in length and 15 Å in diameter, through which the polypeptide extends while still connected to the peptidyl transferase centre^{4,5}. The tunnel is long enough to accommodate a 35-amino-acid-long segment of a nascent polypeptide in an extended conformation⁶. The diameter of the tunnel allows for a helical structure of the peptide, but it is unlikely that significant protein folding can occur beyond helix formation⁵.

The initial folding events inside the ribosomal tunnel, however, are only the first steps in the complex process of folding newly synthesized proteins to their native three-dimensional structure. Whereas some proteins were reported to fold productively in a post-translational mode^{7,8}, several studies demonstrated co-translational folding in the cytosol of prokaryotic and eukaryotic cells⁹. Sequential co-translational folding of domains could be particularly advantageous for the production of multi-domain proteins because it limits possible unproductive inter- and intramolecular interactions during the early folding steps.

The cellular strategy to promote both co- and post-translational folding involves a large arsenal of molecular chaperones^{1,2}. These proteins are found in all kingdoms and can be divided into two groups. The first, mainly chaperones from the Hsp70/40- and Hsp60/10 families, bind to newly synthesized proteins as soluble components of the cytosol. The second group welcomes nascent chains on the ribosome by binding to both the ribosome and the nascent chain, thereby assisting protein folding during ongoing synthesis. The existence of ribosome-associated chaperones is a highly conserved principle in eukaryotes and prokaryotes, although the involved components differ between species. The mechanism of how these ribosome-associated chaperones support the folding of nascent polypeptides is still unclear.

The bacterial ribosome-associated chaperone trigger factor binds to nascent chains and associates with ribosomes in a 1:1 stoichiometry through interactions of its N-terminal domain with the ribosomal protein L23, which is located next to the peptide tunnel exit¹⁰. This association is crucial for its interaction with nascent polypeptides and its *in vivo* function¹⁰. Although trigger factor is not essential in *E. coli*, the combined absence of trigger factor and the cytosolic DnaK chaperone causes cell death above 30 °C and is accompanied by the massive aggregation of more than 340 different newly synthesized proteins^{11,12}. Trigger factor was biochemically defined as a three-domain protein composed of an N-terminal ribosome-binding domain, a central peptidyl-prolyl *cis/trans* isomerase (PPIase) domain and a large C-terminal portion of unknown function¹³. The atomic structures of the isolated N-terminal and PPIase domains have been solved, however, without providing mechanistic insights into the action of trigger factor in protein folding^{14,15}.

Here, we report the crystal structures of *E. coli* trigger factor and

the N-terminal trigger factor fragment bound to *H. marismortui* 50S. Together, these results outline the structure of the trigger factor-ribosome complex and the molecular environment for the folding of newly synthesized polypeptides, and provide a structural basis for the understanding of this fundamental cellular process.

Full-length trigger factor crystallizes in space group $P2_1$ with two molecules in the asymmetric unit. Phases were calculated from a multiple anomalous dispersion experiment on selenomethionine-substituted trigger factor with data extending to 2.7 Å. These data were also used for structure refinement (Supplementary Table 1a). The conformation of both molecules in the asymmetric unit is identical apart from local flexibility in loop regions (Supplementary Fig. S2). Trigger factor folds into a unique shape resembling a crouching dragon with overall dimensions of $122 \times 59 \times 63$ Å³, much larger than expected for this 48-kDa protein (Fig. 1a).

The N-terminal domain of trigger factor ('tail') is exclusively responsible for the interactions with the non-translating ribosome^{10,13}. Curiously, the second domain in sequence harbouring the PPIase activity is located on the opposite side of the molecule ('head'). It is connected to the N-terminal domain by means of a long linker extending along the 'back' of trigger factor. The C-terminal domain is mostly α -helical. It contributes to the formation of the 'back' of trigger factor and adds two extended 'arms' in the core of the protein. This domain is structurally similar to the chaperone domain of SurA, a protein that participates in the folding of outer membrane proteins in Gram-negative bacteria, although there is no sequence homology¹⁶. A large cradle is formed between the N-terminal ribosome-binding 'tail' and the C-terminal 'arms' of trigger factor. The analysis of crystal packing reveals the general peptide binding capacity of this 'chaperone' region. Trigger factor forms an extensive network of intermolecular contacts burying 3,500 Å² of surface area through intertwining of the 'arms' and the inner portion of the ribosome-binding domain. In one of the observed interactions, a neighbouring molecule inserts an α -helix and a β -strand between the 'arms' of the C-terminal domain, which is the narrowest part of the cradle (Supplementary Fig. S3). Even larger structural features could be accommodated between the ribosome-binding domain and the C-terminal 'arms'.

To understand the function of trigger factor as a ribosome-associated chaperone, we co-crystallized the ribosome-binding domain of trigger factor and the large ribosomal subunit (50S). Owing to the suitability of the *H. marismortui* 50S for crystallization we have chosen to form a heterologous complex between *E. coli* trigger factor and *H. marismortui* 50S. We could show conservation of the binding site between the bacterial chaperone and the archaeal ribosome by demonstrating that trigger factor binds with a 1:1 stoichiometry and can be crosslinked to ribosomal protein L23, in an identical way to that observed for the homologous complex (Supplementary Fig. S4). *H. marismortui* 50S was co-crystallized with the 144-amino-acid N-terminal ribosome-binding fragment of trigger factor yielding crystals that are isomorphous to crystals of isolated 50S. This strategy was best suited for obtaining structural data on the trigger factor-ribosome complex because full-length trigger factor interferes with the 50S crystal packing, whereas its connection to the ribosome is too delicate to support alternative packing modes.

Data from co-crystals were collected to 3.5 Å resolution (Supplementary Table 1b) and Fourier difference maps revealed electron density for the ribosome-binding region of trigger factor, which contains the 'signature motif' (₄₃GFRxGxxP₅₀) and two surrounding α -helices (amino acids 20–39 and 50–59). From its anchor point on the ribosome, the trigger factor fragment is visible up to a distance of 20 Å above the 50S surface before the density fades out due to increasing temperature factors. Although flexible in the uncomplexed form, one of the previously observed conformations of the ribosome-binding loop (Protein Data Bank: 1OMS_B; ref. 14)

can be unambiguously placed into the electron density using a kink in one helix and bulky residues as points of reference (Fig. 1b). Starting from the structure of 50S (Protein Data Bank:1JJ2) and the fitted fragment of trigger factor (amino acids 25–59), rigid body fitting, positional and temperature factor optimization and manual rebuilding in the attachment region yielded a model with an $R/R_{\text{free}} = 0.192/0.268$.

Trigger factor binds the ribosome at the triple junction between domain III of 23S ribosomal RNA and proteins L23 and L29, proximal to the exit of the ribosomal nascent chain tunnel (Fig. 1c). Overall, 1,227 Å² of surface area is buried upon complex formation, which is divided into 423 Å², 622 Å² and 182 Å² for the contact of trigger factor with 23S rRNA, L23 and L29, respectively. Because the proteins at the interface form a ridge, only residues at the tip of the ribosome-binding loop around Arg 45 of trigger factor extend far enough to reach the rRNA. Hydrophobic interactions

with L23 involve three residues of trigger factor (Phe 44, Pro 50 and Ile 53) surrounding Met 16 on L23 (Fig. 1d). This residue is replaced by phenylalanine in the *E. coli* L23 protein, which would result in even tighter hydrophobic interactions. The most important interaction is mediated by Glu 13 on L23, which was shown to be essential for trigger factor binding to 50S¹⁰. This residue has an altered side-chain conformation in the complex structure and, with two hydrogen bonds, positions the side chain of Arg 45 of trigger factor for hydrophobic stacking with the unpaired RNA base, and for salt interactions with the backbone phosphate of A1501 in 50S (*E. coli*: A1392) (Fig. 1d). The binding region of L23 and the secondary structure and sequence of rRNA around A1501 are highly conserved in archaea and bacteria (Supplementary Fig. S5a, b). In a second and minor interaction site the C-terminal residues Val 84 and Phe 85 of L23, which were not visible in the isolated 50S crystal structure, form weak hydrophobic contacts with Ile 53 and

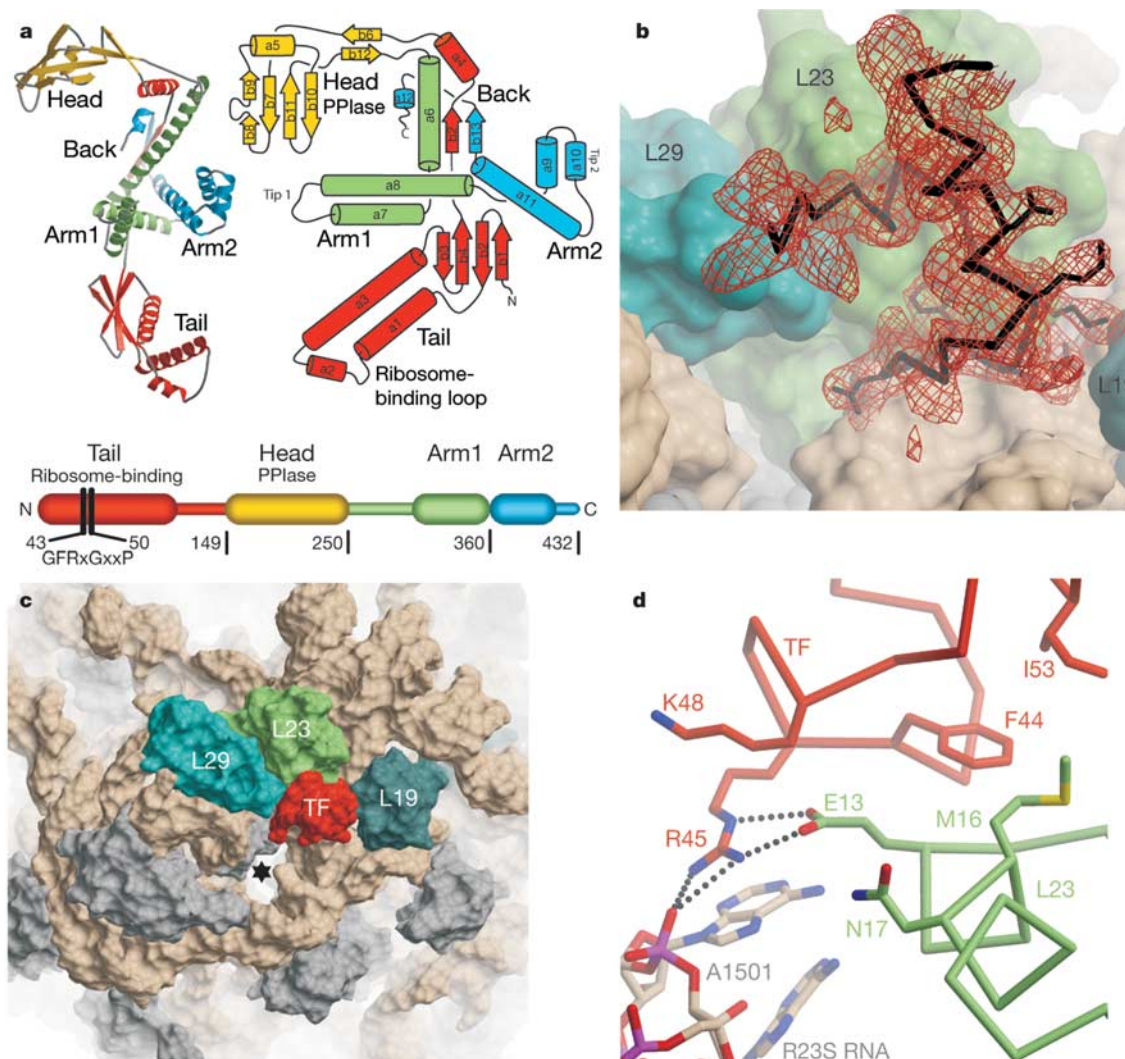


Figure 1 Structure of *E. coli* trigger factor and its N-terminal domain bound to the 50S ribosomal subunit. **a**, Trigger factor adopts an extended fold. Left: ribbon diagram of the trigger factor fold. Right: schematic representation of the domain organization. Bottom: domain arrangement in sequence space. Positions of the ribosome-binding trigger factor signature (residues 43–50) and domain borders are indicated. In all parts the ribosome binding ‘tail’ is shown in red, the PPlase ‘head’ in yellow and ‘arm’ 1 and ‘arm’ 2 in green and blue, respectively. **b**, Trigger factor fragment 1–144 bound to the 50S ribosomal subunit: simulated annealed omit map around residues 26–59 of the bound trigger factor 1–144 fragment shown together with a surface representation of 50S; key interacting

proteins are labelled. For clarity, only selected side chains are shown. **c**, Same site as in **b** viewed from a distance; trigger factor fragment is shown as a red surface representation and the peptide exit tunnel is denoted with an asterisk. In **b** and **c** rRNA is coloured biege; non-interacting proteins, grey; L29, turquoise; L23, green; and L19, bluish-green. **d**, Key interactions between the trigger factor (TF) fragment (red), L23 (green) and 23S rRNA. Selected residues are shown together with C_α-traces. Hydrogen bonds are indicated only for the interaction between the key contact residues E13 (L23), R45 (trigger factor) and A1501 (R23S).

Asn 52 of trigger factor, respectively. Contacts with ribosomal proteins L29 and L19 are tangential and are unlikely to contribute to the trigger factor binding affinity, in agreement with biochemical data¹⁰.

The structure of full-length trigger factor, together with that of its ribosome-binding fragment in complex with 50S, permits accurate placement of the entire chaperone on the ribosome through superposition of the ribosome-binding region that is visible in both structures. Judging from the increasing temperature factors of the helices in the co-crystallized fragment, full-length trigger factor bound to the ribosome could swing, on average, by 10° in all directions around its attachment point. On the basis of the superposition of the two full-length trigger factor molecules in the crystal (Supplementary Fig. S2) and the presence of a well-defined electron density for the interdomain connections, we have no indications for potential domain rearrangements. The peripheral mode of interaction of trigger factor with 50S is also unlikely to induce any remote domain motions in trigger factor. However, the presence of additional conformational states of ribosome-bound trigger factor cannot be ruled out completely.

Trigger factor hunches over the polypeptide exit of the ribosome and extends its sticky, hydrophobic inner face towards the area of the nascent polypeptide exit (Fig. 2a–c). This evokes a model in which nascent chains initially interact with the cradle of the trigger factor surrounded by helices belonging to the ribosome-binding domain and the ‘arms’ of the chaperone. Although trigger factor hunches over the exit and provides a significantly shielded environment for the newly synthesized peptides, no region of the chaperone other than the ‘tail’ contacts the ribosome (Fig. 2a). The closest

approach from the ribosomal surface to the tips of the C-terminal ‘arms’ is around 10 Å whereas the distance to the bottom of the cradle is 40 Å (Fig. 3b). The charge distribution of trigger factor shows an accumulation of basic residues at the ribosome-binding ‘tail’ and of hydrophobic residues both in parts of the N-terminal domain and the C-terminal ‘arms’, supposed to be the contact regions of the cradle for emerging nascent polypeptides (Fig. 3a). These findings agree with biochemical data that demonstrate that trigger factor lacking the PPIase domain still binds peptides enriched in hydrophobic residues¹⁷.

Although the role of trigger factor in protein folding is firmly established, it has been difficult to study its detailed mechanism of action. The role of its C-terminal domain was particularly puzzling because, as an isolated fragment, it is structurally unstable and without assignable function^{13,17}. The atomic structure of trigger factor now provides an explanation for these difficulties. The ‘back’ of the C-terminal domain is critically stabilized by residues 111–133, which connect the N-terminal and the PPIase domain, and therefore may not preserve its conformation in the absence of this connection. Furthermore, from the structure it is not surprising that this domain shows little sequence conservation. It is likely that only the overall shape and the surface distribution of hydrophobicity are critical for substrate binding, and that, due to the diverse nature of nascent peptides, there is little selective pressure to maintain exact positioning of particular side chains.

The C-terminal and the N-terminal domain together form a cradle whose inner surface appears to be tailored to interact with unfolded proteins. Moreover, this cradle is optimally positioned at the exit site of the ribosomal tunnel for the capture of nascent

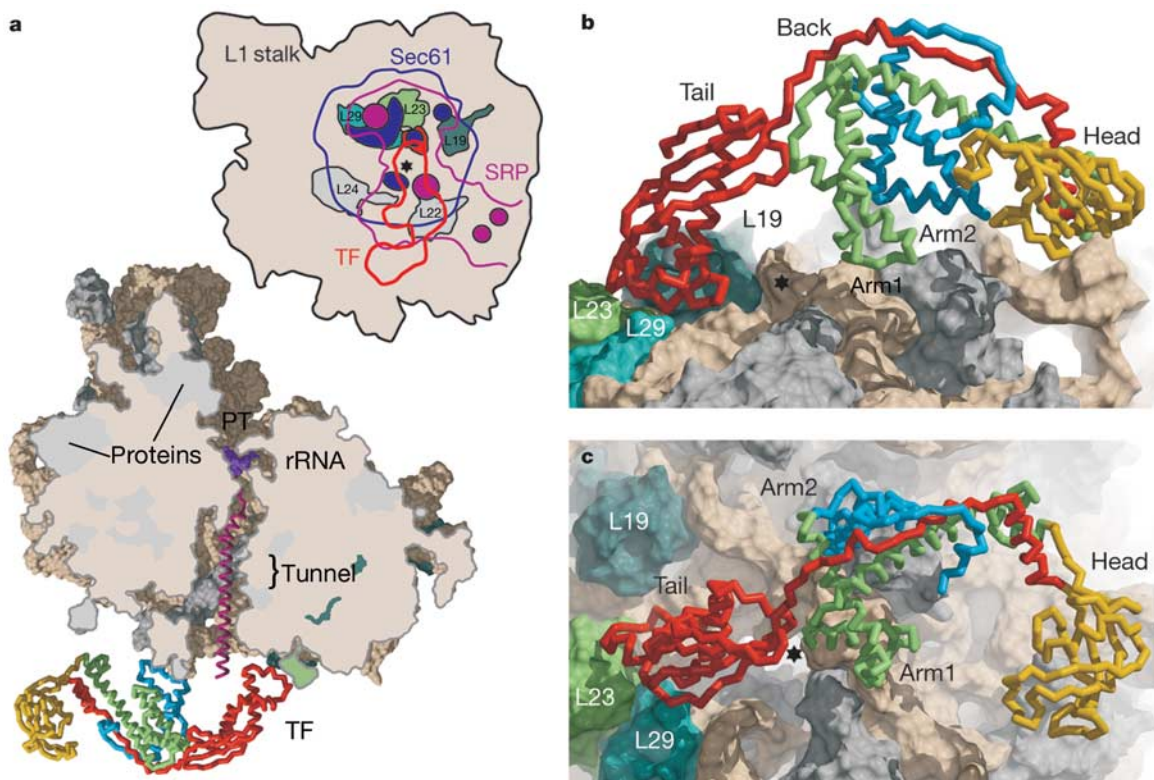


Figure 2 Structure of the trigger factor bound to the 50S ribosomal subunit. **a**, Overview of the trigger factor 50S complex. Full-length trigger factor positioned by superimposition onto the ribosome-bound fragment trigger factor 1–144 is shown as C α -trace together with a slice of 50S along the peptide exit tunnel (for clarity, further cavities peripheral to the tunnel are not shown) with a modelled nascent chain in magenta, extending from the peptidyl transferase centre (PT). Colouring is as in Fig. 1. Inset: schematic footprints of

Sec61p, SRP and the trigger factor on the ribosome on the basis of crystallographic and electron microscopic data. Binding sites for Sec61p (blue), SRP (magenta) and the trigger factor (red) are represented as filled areas, and projections of the molecules are shown as outlines in the same colours. Positions of selected ribosomal proteins are indicated in the same colours as in Fig. 1. **b**, **c**, Close-up side (**b**) and top (**c**) views of the complex shown in **a** without a nascent peptide.

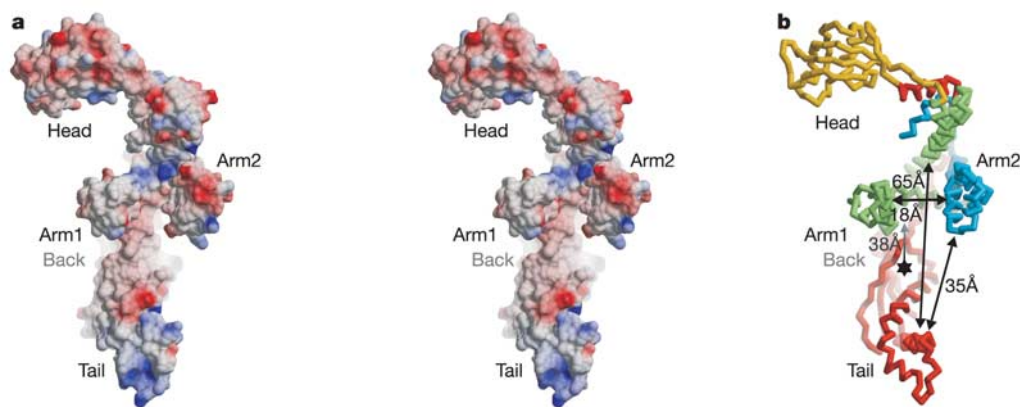


Figure 3 Trigger factor exposes a hydrophobic cradle to the nascent chain. **a**, Solvent accessible surface of trigger factor (in stereo view), coloured by electrostatic potential

(blue, positive; red, negative). **b**, C_{α} -trace representation of trigger factor with approximate dimensions of the cradle indicated; coloured according to Fig. 1.

polypeptides. Its role as the major substrate-binding site of trigger factor is further indicated by several pieces of evidence: (1) there is significant hydrophobicity over the entire inner surface of the cradle (Fig. 3a); (2) the fused N-terminal and C-terminal domains provide trigger factor with almost wild-type-like chaperone activity *in vivo* and *in vitro*¹⁷ as well as peptide binding activity, even in the absence of the PPIase domain (see below); (3) the cradle is involved in forming an extensive network of crystallographic intermolecular contacts as described above; and (4) the structurally similar domain of the SurA chaperone has substrate-binding properties¹⁹.

Recently, it has been shown that PPIase activity is not essential for nascent chain binding and chaperone activity of trigger factor¹⁸. In agreement with these findings, the PPIase domain is peripheral to the main substrate-binding cradle of trigger factor and to the exit of the ribosomal tunnel and is tethered to the body of trigger factor by means of a double linker. As a consequence, the PPIase activity will be limited to regions above the 'arms' and behind the 'back' of trigger factor where partially folded proteins will appear only after they have escaped from the trigger factor cradle. The PPIase domain will therefore interact with nascent chains at later stages of their synthesis, restricting its catalytic activity to selected sites that remain accessible at this folding stage.

Dimerization of trigger factor in solution has been observed *in vitro*, raising possible implications for the mechanism of its action^{17,20}. Crystallographic data presented here show no evidence of a perfect dimer formation. Trigger factor molecules in the crystal form a network of contacts mainly through residues implicated in formation of the substrate-binding cradle and thus may mimic the interactions between trigger factor and its nascent chain substrates as discussed before (Supplementary Fig. S3). These interactions are not possible when trigger factor binds the ribosome because the hydrophobic contact surface orients towards the ribosomal exit tunnel, consistent with the observation that it binds the ribosome as a monomer^{20,21}. Dimerization as observed in solution may help to prevent aggregation and unwanted substrate interactions of trigger factor in the cytosol²⁰.

Trigger factor has inherent affinity for ribosomes, which is increased when they display nascent peptides²¹. The weak affinity of trigger factor to the ribosome ($K_d = 1.2 \mu M$ ^{20,22}) and its small anchor point may be ideally suited for its role in protein folding. By maintaining a precise balance of binding affinities, trigger factor dissociates from vacant ribosomes within seconds²², but may become stabilized by hydrophobic interactions with an emerging, unfolded nascent peptide. The broad distribution of hydrophobicity within the trigger factor cradle may allow extensive but transient contacts of trigger factor with the emerging chain,

consistent with the observation that the association of trigger factor with the translating ribosomes becomes salt insensitive^{1,21}. Once a domain folds co-translationally, it buries hydrophobic residues, detaches from trigger factor and destabilizes the inter-

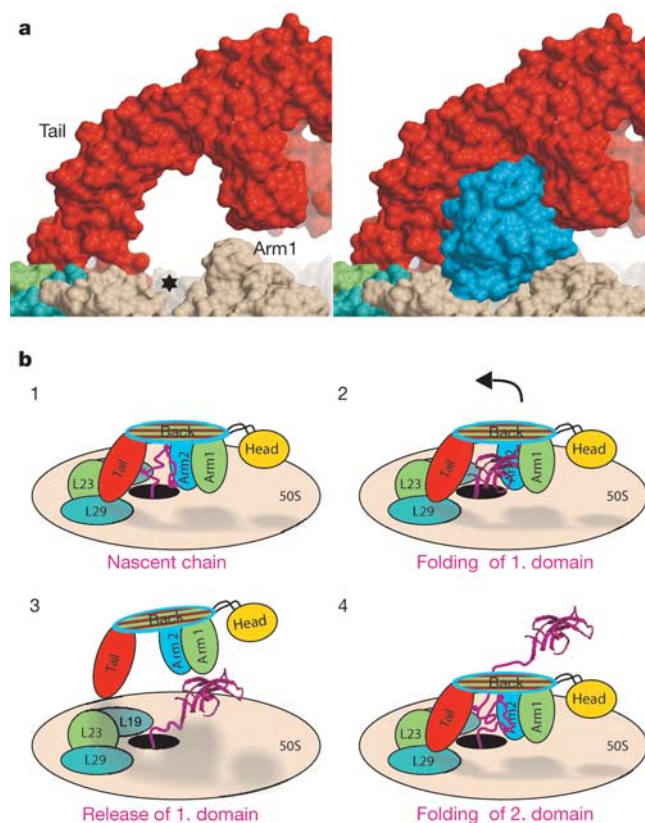


Figure 4 A model for the interaction of trigger factor with nascent chains. **a**, The trigger factor cradle could accommodate entire protein domains. Left: view through the arch-like structure delineated by the tail and arms of trigger factor (red) covering the peptide exit tunnel, indicated by an asterisk. Right: identical view to the lefthand figure; the size of the arch can accommodate an entire molecule of lysozyme (blue), manually fitted into this region for illustration. **b**, Schematic representation of the proposed mechanism of trigger factor action and its possible role in facilitating domain-wise co-translational protein folding. Colours are as in Fig. 2. Initially (1) the trigger factor is bound to an unfolded nascent chain. Upon folding of this domain (2), contacts between the trigger factor and the newly synthesized peptide are weakened and trigger factor dissociates from the ribosome (3). The trigger factor re-associates with the ribosome when the next stretch of newly synthesized, unfolded polypeptide becomes exposed (4).

action of trigger factor with the ribosome–nascent chain complex (Fig. 4b).

We conclude that trigger factor might promote co-translational folding of domains by providing a shielded environment in which the folding is initially postponed through hydrophobic contacts, but then allowed to proceed, perhaps in a self-promoted fashion, when sufficient sequence information is available for the generation of a folded core (Fig. 4b). The chaperone would then dissociate from the ribosome and rebind once a significant new portion of the unfolded polypeptide becomes exposed at the tunnel exit. The folding space formed between trigger factor and the ribosome is sufficiently large to accommodate a domain of the size of the 14-kDa protein lysozyme (Fig. 4a). A different option, which we consider less likely on the basis of available biochemical evidence, is that trigger factor remains bound during the synthesis of the entire polypeptide, even when multi-domain proteins are translated. In this case, the co-translationally folded domain could escape through the relatively large gaps between the ribosome and the chaperone arch on either side.

Trigger factor is just one of the macromolecular factors that bind to ribosome–nascent chain complexes. The signal recognition particle (SRP) targets ribosomes displaying signal sequences to membranes where the newly synthesized proteins are co-translationally translocated²³. At the membranes, the ribosome–nascent chain complexes are bound to the translocon, a protein-conducting channel. Both SRP and the translocon have now been visualized on the eukaryotic ribosome by electron microscopy^{24,25}. Because protein folding and targeting require precise exchange between all these factors it is interesting to compare their binding sites on the ribosome. Both SRP and the eukaryotic protein-conducting channel, Sec61, use eukaryotic homologues of ribosomal proteins L23 and L29 as major attachment sites, analogous to what is reported here for trigger factor. These ribosomal proteins have also been crosslinked to the protein component of SRP (SRP54)²⁶. It is tempting to speculate that all factors transiently associated with ribosome–nascent chain complexes will use a similar mode of interaction characterized by low affinity, loop-mediated binding to the ribosome augmented by contacts to the nascent chain. Interestingly, one of the attachment sites for Sec61 coincides with those of trigger factor (Fig. 2a).

Here we have shown that trigger factor, together with the ribosome, forms a well-defined protective cage of sufficient size to accommodate a folded protein domain at the peptide tunnel exit site. This is in contrast to the previous assumption that the 48-kDa trigger factor can only bind small stretches of unfolded peptides but not provide a shielded folding space. The two modes of action may not be mutually exclusive and may depend on the nature of the nascent polypeptide. The structural information presented here will serve as a basis for functional experiments aimed at addressing the mechanism of protein folding facilitated by ribosome-associated chaperones. □

Methods

Crystallization of trigger factor and trigger factor–50S complexes

Full-length trigger factor and its N-terminal fragment consisting of the first 144 amino acids were expressed and purified as previously described¹⁸. We obtained crystals of the full-length trigger factor in 22–26% (w/v) polyethylene glycol 4,000 with 14% (v/v) 1,4-butanediol in 0.1 M MES (2-N-morpholino-ethanesulphonic acid) at pH 7.0–7.5. Crystals grew to a maximum size of $0.2 \times 0.3 \times 1.0$ mm³ in four to six weeks at 4 °C and belong to space group $P2_1$ ($a = 100.2$ Å, $b = 47.4$ Å, $c = 114.8$ Å, $\beta = 113.6^\circ$) with two molecules in the asymmetric unit. Selenomethionine-labelled protein was produced with standard procedures and crystallized under similar conditions. The N-terminal fragment of trigger factor was used at tenfold molar excess over 50S for co-crystallization experiments. Crystals were obtained under conditions previously described for 50S crystallization⁴ and belong to space group $C222_1$ ($a = 211$ Å, $b = 299$ Å, $c = 574$ Å) with one molecule in the asymmetric unit.

Data collection

Full-length trigger factor crystals were stabilized by stepwise addition of 2-methyl-2,4-pentane diol (MPD) to yield a final concentration of 25% (v/v). Crystals were flash frozen in liquid propane. All measurements were done at the Swiss Light Source beamline X06SA

at the Paul Scherrer Institute at 100 K. A native data set to 2.8 Å resolution and a selenomethionine derivative data set at inflection point, peak and remote wavelength to 2.7 Å were collected. The latter was used for multiwavelength anomalous dispersion phasing (Supplementary Table 1a) and, owing to the large non-isomorphism between crystals, also for refinement.

Crystals of 50S in complex with the ribosome-binding domain of trigger factor were stabilized as described and flash frozen in liquid propane⁴. Complete data sets were collected at the Swiss Norwegian Beamline at ESRF and the Swiss Light Source to 3.5 Å resolution (Supplementary Table 1b). All data were integrated and scaled using HKL²⁷.

Structure determination

All phasing and refinement was done using the program CNS²⁸. For the full-length trigger factor data, an initial set of heavy atom sites was identified in the anomalous difference Patterson maps by automatic correlation coefficient search procedures. Additional sites were found in difference Fourier maps after improving the phases by solvent flipping. In total, 19 selenium sites were used for phasing. Maximum likelihood refinement with amplitudes and experimental phase probability distribution was used during model building. The structure was refined to 2.7 Å resolution with $R/R_{\text{free}} = 0.241/0.324$ with an average B-value of 74 Å². Improved maps were calculated as prime-and-switch composite omit maps using the program RESOLVE²⁹. The N-terminal ribosome binding domain of the trigger factor in complex with the 50S was located in the $F_o - F_c$ difference Fourier maps. The structure of the complex was refined to 3.5 Å resolution with $R/R_{\text{free}} = 0.192/0.268$. All models were built using the program O³⁰. For details of figure preparation see Supplementary Information.

In vitro crosslinking

1 µM purified 50S ribosomal *H. marismortui* subunits were incubated with 2 µM mutant trigger factor V49C coupled to the UV-inducible crosslinker BPIA (benzophenone-4-iodoacetamide, Molecular Probes, Inc.) for 30 min at 30 °C in the presence of 0.6 M NaCl as described¹⁰. Samples were irradiated with UV light on ice for 5 min (365 nm, 100 W; Ultraviolet Products, Model B-100AP) at a distance of 5 cm. Ribosomal complexes were separated by centrifugation through sucrose cushions¹³. Protein identification was done by in-gel-digest of the crosslinking products with trypsin followed by mass spectrometry¹⁰.

Received 25 May; accepted 27 July 2004; doi:10.1038/nature02899.

Published online 29 August 2004.

1. Bukau, B., Deuerling, E., Pfund, C. & Craig, E. A. Getting newly synthesized proteins into shape. *Cell* **101**, 119–122 (2000).
2. Hartl, F. U. & Hayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**, 1852–1858 (2002).
3. Ramakrishnan, V. Ribosome structure and the mechanism of translation. *Cell* **108**, 557–572 (2002).
4. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
5. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
6. Hardesty, B. & Kramer, G. Folding of a nascent peptide on the ribosome. *Prog. Nucleic Acid Res. Mol. Biol.* **66**, 41–66 (2001).
7. Agashe, V. R. *et al.* Function of trigger factor and DnaK in multidomain protein folding: increase in yield at the expense of folding speed. *Cell* **117**, 199–209 (2004).
8. Netzer, W. J. & Hartl, F. U. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388**, 343–349 (1997).
9. Nicola, A. V., Chen, W. & Helenius, A. Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nature Cell Biol.* **1**, 341–345 (1999).
10. Kramer, G. *et al.* L23 protein functions as a chaperone docking site on the ribosome. *Nature* **419**, 171–174 (2002).
11. Deuerling, E., Schulze-Specking, A., Tomoyasu, T., Mogk, A. & Bukau, B. Trigger factor and DnaK cooperate in folding of newly synthesized proteins. *Nature* **400**, 693–696 (1999).
12. Teter, S. A. *et al.* Polypeptide flux through bacterial Hsp70: DnaK cooperates with Trigger Factor in chaperoning nascent chains. *Cell* **97**, 755–765 (1999).
13. Hestekamp, T., Deuerling, E. & Bukau, B. The amino-terminal 118 amino acids of *Escherichia coli* trigger factor constitute a domain that is necessary and sufficient for binding to ribosomes. *J. Biol. Chem.* **272**, 21865–21871 (1997).
14. Kristensen, O. & Gajhede, M. Chaperone binding at the ribosomal exit tunnel. *Structure (Camb.)* **11**, 1547–1556 (2003).
15. Voght, M. *et al.* NMR solution structure and dynamics of the peptidyl-prolyl cis-trans isomerase domain of the trigger factor from *Mycoplasma genitalium* compared to FK506-binding protein. *J. Mol. Biol.* **318**, 1097–1115 (2002).
16. Bitto, E. & McKay, D. B. The periplasmic molecular chaperone protein SurA binds a peptide motif that is characteristic of integral outer membrane proteins. *J. Biol. Chem.* **278**, 49316–49322 (2003).
17. Kramer, G. *et al.* Functional dissection of *Escherichia coli* trigger factor: unraveling the function of individual domains. *J. Bacteriol.* **186**, 3777–3784 (2004).
18. Kramer, G. *et al.* Trigger factor peptidyl-prolyl cis/trans isomerase activity is not essential for the folding of cytosolic proteins in *Escherichia coli*. *J. Biol. Chem.* **279**, 14165–14170 (2004).
19. Behrens, S., Maier, R., de Cock, H., Schmid, F. X. & Gross, C. A. The SurA periplasmic PPIase lacking its parvulin domains functions *in vivo* and has chaperone activity. *EMBO J.* **20**, 285–294 (2001).
20. Patzelt, H. *et al.* Three-state equilibrium of *Escherichia coli* trigger factor. *Biol. Chem.* **383**, 1611–1619 (2002).
21. Hestekamp, T., Hauser, S., Lütcke, H. & Bukau, B. *Escherichia coli* trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. *Proc. Natl Acad. Sci. USA* **93**, 4437–4441 (1996).
22. Maier, R., Eckert, B., Scholz, C., Lilie, H. & Schmid, F. X. Interaction of trigger factor with the ribosome. *J. Mol. Biol.* **326**, 585–592 (2003).

23. Schatz, G. & Dobberstein, B. Common principles of protein translocation across membranes. *Science* **271**, 1519–1526 (1996).
24. Beckmann, R. *et al.* Alignment of conduits for the nascent polypeptide chain in the ribosome–Sec61 complex. *Science* **278**, 2123–2126 (1997).
25. Halic, M. *et al.* Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature* **427**, 808–814 (2004).
26. Pool, M. R., Stumm, J., Fulga, T. A., Sinning, I. & Dobberstein, B. Distinct modes of signal recognition particle interaction with the ribosome. *Science* **297**, 1345–1348 (2002).
27. Otwinowski, Z. M. W. in *Macromolecular Crystallography Part A* Vol. 276 (ed. Carter, C. W. Jr, S.R.M.) 307–326 (Academic, San Diego, 1997).
28. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
29. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 849–861 (1999).
30. Jones, T. A., Bergdoll, M. & Kjeldgaard, O. in *Crystallographic and Modeling Methods in Molecular Design* (eds Bugg, C. & Ealick, S.) 189–195 (Springer-Verlag, New York, 1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements Data collection was performed at the Swiss Light Source, Paul Scherrer Institut, Villigen and at the Swiss Norwegian Beamline (ESRF, Grenoble). We are grateful to C. Schulze-Bries, T. Tomizaki and A. Wagner at the SLS as well as P. Pattison and S. Capelli at the SNBL whose outstanding efforts have made these experiments possible. We also thank our colleagues S. Antolic, M. Steiner and members of the Ban laboratory for help in ribosome preparations and suggestions; and members of the Bukau laboratory for discussions. This work was supported by the Swiss National Science Foundation (SNSF), the NCCR Structural Biology program of the SNSF, an ETH research grant, grants of the Deutsche Forschungsgemeinschaft to B.B. and E.D. and a Young Investigator grant from the Human Frontier Science Program to N.B. and E.D.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.B. (ban@mol.biol.ethz.ch) or E.D. (e.deurling@zmbh.uni-heidelberg.de). Structure coordinates have been deposited in the Protein Data Bank under the accession codes 1w26 and 1w2b.

The silent treatment

Biotech firms are vying to harness the potential of RNA interference. But will its impact be in finding new disease targets, or in RNA-based drugs? Julie Clayton investigates.

The hype surrounding RNA interference (RNAi) 'gene-silencing' technology has both academic labs and biotech companies firmly in its grip. RNAi is lauded as a powerful approach to gene control, with the promise of revolutionizing basic research and providing treatments for intractable conditions such as cancer and neurodegenerative disease.

Far from being an established tool, RNAi is still a maturing technology with some way to go before its true potential is known. Many in the field are divided over whether its greatest contribution will lie in the identification and validation of disease targets — for new drugs ranging from small-molecule inhibitors to antibodies — or in the less certain development of RNAi itself as a therapeutic.

Until relatively recently, RNAi was the preserve of those studying plants and invertebrates, notably the nematode *Caenorhabditis elegans*. But three years ago Tom Tuschl and his colleagues, then at the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany, showed that the technology could be used reliably to silence gene expression in mammalian cells (S. M. Elbashir *et al.* *Nature* **411**, 494–498; 2001). "It took everybody by surprise", says Nassim Usman, senior

vice-president of Sirna Therapeutics in Boulder, Colorado, one of the firms aiming to develop RNAi as a therapy.

RNAi relies on the introduction of double-stranded RNA molecules into cells to block the translation of messenger RNAs into protein. The double-stranded RNA, which must be partly identical in sequence to the gene to be inhibited, is cleaved into shorter fragments in the cell by an enzyme called Dicer, and the 'sense' strand degraded. The remaining 'antisense' strand then becomes incorporated into a protein complex called RNA-induced silencing complex. The activity of this complex is directed by the RNA towards the target messenger RNA, which is degraded, thus effectively 'silencing' the gene from which it was transcribed (C. D. Novina and P. A. Sharp *Nature* **430**, 161–168, 2004; G. J. Hannon *Nature* **418**, 244–251, 2002; and *Nature*, **431**, 337–378, 2004).

Key to Tuschl's success was the discovery that if the introduced RNAs were short enough — between 21 and 23 nucleotides long — they would not trigger the mammalian cell's defence mechanism against RNA



Dmitry Samarsky: on target for RNAi.

viruses, the production of interferon, which was masking the effect of gene silencing. Attention turned to these short RNAs, called small interfering RNAs (siRNA), which can be synthesized chemically to match the desired target gene.

A slew of studies then showed that RNAi was an impressive tool for manipulating genes in laboratory-cultured mammalian cells,

raising the tantalizing prospect of doing the same in animal models, and possibly even in humans. With impressive speed, life-sciences equipment supply companies have developed kits and reagents for supporting RNAi-based research (see 'RNAi options', below). At the same time, new companies have emerged, hungry for partners, offering large-scale and high-throughput RNAi-based analysis of gene function, for both basic research and for the identification and validation of new drug targets. "RNAi has turned out to be a wonderful target-validation tool", says Dmitry Samarsky, head of business and technology development at California-based Invitrogen.

RNAi OPTIONS

Kits and reagents for making the various types of RNA interference (RNAi) constructs and getting them into cells are now widely available. The simplest options are designed to enable the negatively charged, chemically synthesized small interfering RNA (siRNA) molecules to enter cells more easily. One way is to use electroporation. Alternatively, the siRNA can be encapsulated in lipids or polymers to overcome the repulsive charges of the cell membrane, as in kits supplied by companies such as Ambion in Austin,

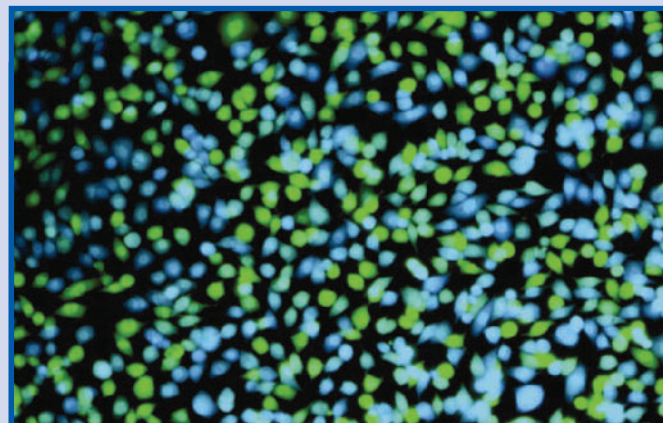
Texas, Invitrogen in Carlsbad, California, and QIAGEN in Venlo, the Netherlands. QIAGEN has developed siRNA sets against specific functional gene families and pathways, such as those leading to apoptosis and cancer. Their Human Druggable Genome siRNA Set targets 5,000 human genes of therapeutic interest. For high-throughput siRNA-based screening, QIAGEN offers an RNAi human/mouse control kit that provides all the reagents necessary for successful start-up. Oligonucleotide producers now also make siRNAs or their DNA equivalents to order, and siRNA manufacturers such as Dharmacon in Dallas, Texas, MWG Biotech in Ebersberg, Germany, and Proligo in Boulder, Colorado, offer free online design services.

Transfected siRNAs achieve significant gene 'knock-down' for some 3–7 days before being naturally degraded. This is usually sufficient for studying the immediate effects of inhibiting gene expression, when screening for drug targets, for example, but is likely not to be sufficient for RNAi-based therapy.

An effect lasting for weeks can be obtained with expression systems based on plasmid or viral vectors carrying DNA versions of the interfering RNA. The DNA is transcribed and the transcript processed into siRNA within the cell. In many of these systems the constructs are expressed as short double-stranded RNA 'hairpins' (shRNA), which are then cleaved by cellular enzymes to produce functional siRNAs. Vectors based on adenoviruses and lentiviruses also enable RNAi to be extended to a wider range of non-dividing primary cells than when transfecting with siRNA or a plasmid. Kits for constructing adenoviral and lentiviral vector RNAi expression systems are supplied by Promega in Madison, Wisconsin, BD Biosciences in San Jose, California, Imgenex in San Diego, California, Ambion and Invitrogen.

J.C.

INVITROGEN



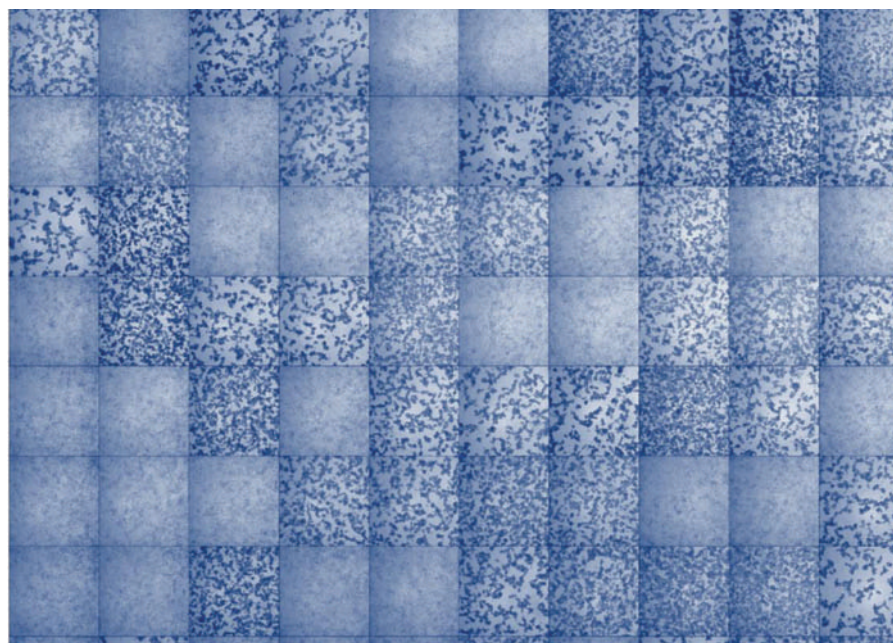
Stop-go: Cells (green) silenced with a viral RNAi vector.

From worms to humans

Cenix BioScience in Dresden, Germany, is one of the new firms that has been formed to seize these opportunities. Chief executive Christophe Echeverri felt nervous about leaving academia as a postdoc to start up the company in 1999, but has never looked back. Cenix is developing RNAi technology for basic research and drug-target validation, and mainly offers *in vitro* screening of cell lines using a library of siRNAs designed against the human genome, and against potentially druggable targets such as kinases (see *Nature* 428, 225–231; 2004). siRNAs “really represent the gold standard reagent with the broadest applicability” compared with vector-expressed short hairpin RNA (shRNA), says Echeverri. The level of gene silencing produced with shRNA is more variable, less reproducible, and does not permit control over the dose achieved inside cells, he says.

“Synthetic siRNA is the preferred choice for high-throughput screening,” agrees Walter Tian, business director at QIAGEN, based in Venlo, the Netherlands. Unlike siRNA, he adds, the vector-based shRNA approach still faces challenges, such as the need to create effective design algorithms and the generation of genome-wide libraries. The need for individual optimization means that the shRNA approach may not be cost-effective for high-volume labs that simply want to screen easily transfected cultured cell lines.

A not-for-profit organization, the Translational Genomics Research Institute in Gaithersburg, Maryland, carries out con-



GALAPAGOS GENOMICS

Scaling up: shRNA knock-down of gene expression in a microtitre plate.

tract screening of cell lines and primary cells for targets related to cancer biology, using QIAGEN's library of 10,000 siRNAs against 5,000 druggable targets, and about 25,000 shRNA vectors. It looks for “points of vulnerability” — new drug targets and ways to enhance the effectiveness of existing drugs, says Spyro Mousses, director of the institute's cancer drug development laboratory.

In contrast to the siRNA approach, Belgium-based biotech company Galapagos

Genomics specializes in high-throughput screening of human primary cells using proprietary adenovirus vectors expressing shRNAs against 4,900 known druggable targets involved in a range of diseases, including rheumatoid arthritis, osteoarthritis, asthma and Alzheimer's disease. The viral vectors have modified protein coats to broaden the target-cell range, and are arrayed in 96- or 384-well plates to give relatively high-throughput screening. Galapagos is also

EXPRESS DELIVERY

At the University of California, Los Angeles, professor of medicine William Pardridge has developed a ‘molecular Trojan horse’ strategy for RNA interference (RNAi) delivery, which is licensed to drug-delivery specialists ArmaGen Technologies in Santa Monica, California. It consists of incorporating either peptides or monoclonal antibodies that target membrane-bound receptors into 85-nanometre-diameter liposomes containing plasmid DNA with the desired siRNA sequence. The lipids are coupled to polyethylene glycol to prevent nonspecific fusion with the cell membranes of non-target cells. This strategy aims both to protect the DNA from degradation by nucleases and to

ensure its sequential uptake by endocytosis across the blood-brain barrier and the cell and nuclear membranes of neurons. The system has been tested in adult mice and rhesus monkeys, and Pardridge hopes to see it in clinical trials against brain cancer “within two years”.

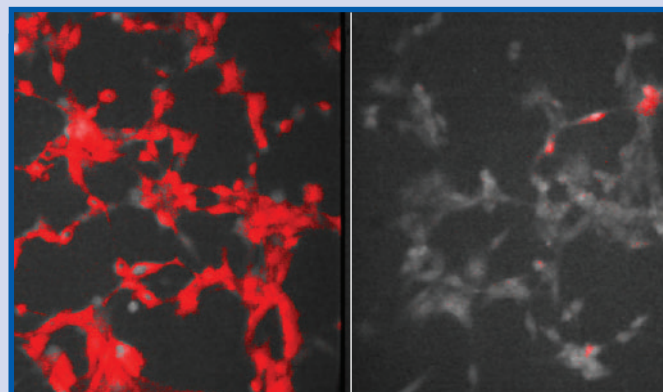
Trojan horses also feature in TargeTran, the delivery system from Intradigm in Rockville, Maryland. TargeTran involves nanoparticles formed by the self-assembly of positively charged polymers with the negatively charged siRNA, encasing and protecting it. Additional ligands are added for targeting to particular tissues after intravenous injection, and combinations of siRNAs can be used to inhibit multiple drug targets.

One of Intradigm's targets is the vascular endothelial growth factor (VEGF) pathway, which is involved in angiogenesis in tumours and other diseases. Martin Woodle, chief scientific officer at Intradigm, hopes to take the VEGF-targeting technology into the clinic, first for cancer next year, and then against age-related macular degeneration, a relatively common cause of blindness, which involves abnormal blood vessel growth in the retina. Alnylam, of Cambridge, Massachusetts, is also developing siRNA therapeutics against this condition.

But Woodle's collaborator Raymond Schiffelers at the Netherlands Cancer Institute in Amsterdam cautions that many ligands targeting multiple receptors are likely to be needed for the clinical application of RNAi against cancer, which in humans “will be far more complex because there are different phases of tumorigenesis where parts of the process are not angiogenic”. Intradigm is also collaborating with researchers in Guang Zhou, Hong Kong and Beijing to produce RNAi-based therapy against the coronavirus that causes severe acute respiratory syndrome (SARS).

J.C.

W. PARDRIDGE/UCLA



Trojan horse: RNAi liposomes silence EGFR function (right).

developing animal models for *in vivo* target validation using shRNA, expressed from an adenovirus vector, says cellular and molecular biology director Helmuth van Es — for example, to target the synovium in a collagen-induced model of arthritis.

High-throughput screening has revealed two bottlenecks in the target-discovery process, according to Sumit Chanda at the Genomics Institute of the Novartis Research Foundation (GNF) in San Diego, California. The first is the need to make RNAi screening in primary cell cultures as fast and efficient as in cell lines. The second is to translate *in vitro* results into animal models.

siRNA delivery to primary cells remains less efficient than in transformed cell lines, and not every primary cell type is amenable to lipid-based transfection. As an alternative, a few companies, including BTX in Holliston, Massachusetts, and Cyto Pulse Sciences in Columbia, Maryland, offer 96-well plate-based electroporation. The GNF is developing high-throughput electroporation that would enable transfection with 20,000 oligonucleotides at a time, says Chanda, but this is not yet commercially available. And attempts to use vectors to deliver individual inhibitory RNAs to 20,000 wells at a time harbour significant technical and safety challenges, Chanda adds, including the need to achieve even titres of virus per well and to conduct all procedures to appropriate biosafety standards.

Down but not out

Several groups are tackling the second bottleneck — *in vivo* validation. Compared with the creation of gene knockouts in mice, which can take up to nine months, RNAi in animals can potentially produce answers within days or weeks, greatly speeding up the number of drug targets that can be validated in a given time. The often incomplete nature of RNAi — usually up to 90% knock-down — means that it may reflect more accurately the situation in human disease, where a disease-causing gene may be operating at suboptimal levels. It also mimics more closely the effect of a small-molecule drug, which usually achieves only incomplete inhibition of its target.

Effective RNAi *in vivo* in animals could mean vast savings for the drug industry, says Martin Woodle of siRNA therapeutics company Intradigm in Rockville, Maryland. Traditionally it can cost from US\$10 million to \$50 million to identify a new small-molecule drug and go to the first tests in animals, making it expensive if drug targets that looked promising *in vitro* fail at this stage. Further savings will occur if the siRNA itself can be developed as the drug.

For research purposes, the GNF is trying to take RNAi into vertebrates with faster generation times than mice, such as zebrafish



Imran Ahmad:
keep options open.

and chicks. But most research groups and companies are concentrating on developing the use of RNAi in mouse models of disease.

Getting it there

The key challenge for achieving effective RNAi *in vivo* is delivery to the desired organ and into the target cells, to ensure specificity and adequate dose. “RNAi will never leave the Petri dish until

we solve the delivery problems,” says William Pardridge, professor of medicine at the University of California, Los Angeles. But it’s good news for the biotech sector that there is unlikely to be a single solution. “I think the delivery problem will be solved — either by delivery locally, into the local blood supply, or by targeting,” says Judy Lieberman, a paediatrician at Harvard University’s CBR Institute for Biomedical Research. “Viral vectors may make more sense for genetic diseases when you want long-term efficacy. There are so many possible indications that some delivery methods may be better than others.”

One approach is to target the RNAi construct specifically at a tissue. Either the siRNA or the DNA encoding it is encased in a liposome or polymer coat, and ligands are incorporated into the coat that bind to cell-surface receptors on the target tissue, which

A RIVAL TO ANTISENSE?

Alongside the optimism over RNA interference (RNAi) comes the inevitable question of whether it will fare better than previous RNA-based technologies — antisense and ribozymes — both in the laboratory and in the clinic.

Many commentators are placing their bets on RNAi because it taps into a pre-existing control system within the cell and has the potential for greater potency, given that the same small interfering RNA (siRNA) molecules can recycle between different copies of messenger RNA. Single-stranded antisense molecules do not exist naturally, and each one acts only once to block translation of messenger RNA before being degraded. Ribozymes do cleave multiple copies of the same messenger RNA, but their potency remains in question, and there is, as yet, no sign of success as potential therapeutics.



Sirna has switched from ribozymes to siRNA.

“RNAi is what antisense never was. It’s robust and reproducible, and it exists in nature,” says Inder Verma of the Salk Institute for Biological Sciences in San Diego, California. RNAi is “much more active and effective than antisense and ribozymes”, according to Martin Woodle of Intradigm in Rockville, Maryland.

Frank Bennett of Isis Pharmaceuticals in Carlsbad, California, disputes this, at least for *in vitro* use, saying that “by and large antisense and RNAi are equally potent in cell culture”. For *in vivo* use, he asserts that RNAi has yet to be fully optimized to justify comparison with antisense. “We’re keeping an open mind regarding the therapeutic potential of RNAi. It’s too early to make predictions,” he says.

Isis is so far the only company to have succeeded in getting an antisense drug licensed — Vitravene for the treatment of skin disease — and has other antisense products in clinical trials. In contrast, the company formerly known as Ribozyme Pharmaceuticals relaunched in 2003 as Sirna Therapeutics after its most promising antisense candidates proved ineffective in clinical trials. The RNAi technology breakthrough occurred at the right time, according to Nassim Usman, senior vice-president at Sirna.

Others point out the additional concerns that antisense molecules can trigger immune reactions, being larger and ‘foreign’ in composition, compared with siRNA, and can be toxic. But Bennett asserts that Isis has modified its antisense products to minimize these potential problems.

NeoPharm in Illinois is maintaining a stake in both approaches. While results from animals show the overall effectiveness of RNAi to be “very much better than with antisense”, according to chief scientific officer Imran Ahmad, and to require a fivefold smaller dose, the company intends to compare its lipid-based delivery system for antisense and RNA side-by-side in the clinic.

J.C.

should then take up the construct by endocytosis (see 'Express delivery', page 601). Another approach is to rely on the physiological likelihood that the construct will be taken up by the desired tissue.

Working with siRNA encased in a cationic polymer, Jianzhu Chen of the Massachusetts Institute of Technology (MIT) Center for Cancer Research has found that in mice the RNA is taken up most efficiently by lung cells following injection into the tail vein. This is probably because the lung contains the first capillary beds traversed by intravenously injected material and is the tissue with the most extensive blood supply.

And NeoPharm in Lake Forest, Illinois, is developing liposome carriers for RNAi-based cancer therapy. By modifying the surface charge of a naturally occurring mitochondrial membrane lipid, cardiolipin, the company has developed a cationic liposome that is less toxic than current positively charged liposomes, according to chief scientific officer Imran Ahmad. To get the RNA to its target, NeoPharm is relying on the preferential uptake of the liposomes as a source of fat by fast-growing tumours.

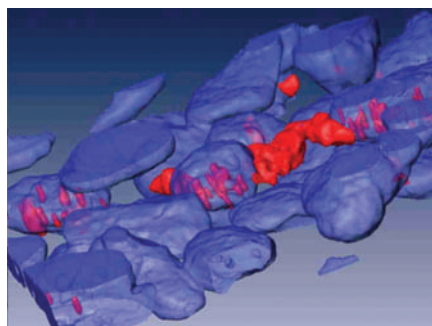
Meanwhile Mirus Bio in Madison, Wisconsin, is developing a new delivery method — initially for gene therapy but eventually also for RNAi-based therapy — that targets skeletal muscle. Tested so far in animals, this involves applying a tourniquet to the leg, to maintain a high local volume of blood, followed by injection into veins in the foot. The company hopes to develop this method to treat muscular dystrophy and ischaemia by gene therapy, says Jim Hagstrom, vice-president of scientific operations.

Viral vectors

Some researchers in the field point out that no one has yet shown chemically synthesized siRNA to work *in vivo* for more than a few weeks, or to be able to target a broad range of tissues. And not everyone is convinced that encasing the nucleic acid in polymers or liposomes is the best way to go.

"It'll be great to have polymers if they work, but I don't think they're efficient enough yet — viruses have learned to do this over millions of years," says Inder Verma, a geneticist at the Salk Institute for Biological Studies in San Diego, who is developing lentivirus-based expression systems for long-term RNAi *in vivo*. The vectors contain DNA inserts linked to a promoter recognized by RNA polymerase III, and are designed to express shRNA inside fertilized mouse eggs, creating 'knock-down' transgenic mice in just three to four weeks, with the effect lasting over many generations.

Unlike plasmids, viral vectors can introduce siRNAs into non-dividing cells such as neurons, and the lentivirus vectors have the advantage over adenovirus of being able



Viral shRNA transcripts (red) being exported from cell nuclei (blue).

to introduce RNAs into blood and bone-marrow cells. Luk Van Parijs and colleagues at MIT have successfully induced RNAi in haematopoietic stem cells using lentivirus vectors.

Both Verma and Van Parijs advocate the use of lentivirus vectors for pinpointing genes involved in complex and chronic diseases such as Alzheimer's disease and type 1 diabetes, and for countering their effects.

According to Verma, many viral expression systems have disadvantages *in vivo* because they cannot be targeted to individual tissues. Instead, they have to be injected locally, or cells need to be removed from the body, treated *ex vivo* and replaced. He also cautions that, as with siRNA, vector-expressed shRNA has the potential to produce non-targeted effects, which could lead to side effects.

Different vectors will probably have to be developed for different tissues. Beverly Davidson, associate director at the University of Iowa Center for Gene Therapy, turned to adenovirus-associated virus (AAV) vectors in her work on a mouse model of spinocerebellar ataxia after finding it "difficult to get siRNA into tissues" *in vivo*, she says. The treatment entails direct injection of the AAV vector into the brain. "AAV is good at transducing in the brain, which is a well-localized organ and does not permit the virus to spread elsewhere in the body," notes Verma. AAV also has the advantage that it has already been through clinical trials for gene therapy and can be produced to a thousand-fold higher concentration than lentiviruses. Davidson sees the prospect for developing this into human therapy as "very exciting", but insists on a conservative estimate on the timeline. "If we could overcome the hurdles, then it would be reasonable to consider human trials within five years," she says.

Meanwhile, John Rossi of the City of Hope hospital in Duarte, California, is developing lentivirus-based shRNA therapy against HIV. The lentiviruses will contain DNA encoding shRNA against the HIV *tat* and *rev* genes, which are expressed early in HIV's life cycle. Stopping the virus in its tracks so early should also minimize the risk

of resistance arising. Rossi hopes to get permission to test the system in AIDS patients undergoing bone-marrow transplants for AIDS-related lymphoma.

Keeping safe

Safety is paramount when it comes to using viral vectors in humans. The current generation of lentivirus vectors are "very safe", according to Rossi, as judged by animal studies and a current clinical trial for HIV using a lentiviral vector to deliver an antisense construct, conducted by VIRxSYS of Gaithersburg, Maryland. Benitec, a company in Queensland, Australia, that specializes in vector-based RNAi, may commercialize Rossi's lentiviral system if it proves successful. The firm is also supporting work on an AAV-based RNAi against hepatitis C virus (HCV) originating in the work of Mark Kay of Stanford University in California. HCV is an ideal target for RNAi, according to Alexander Kolykhalov, scientific director at Benitec, as it is an RNA virus and is constantly available in the cytoplasm.

AAV has only shown minor side effects in gene therapy in dogs, and in a limited trial of gene therapy for muscular dystrophy in humans, and Kay claims that these problems are "technically solvable". He also plans to use a different AAV serotype in his work. But he admits that "it's impossible to predict whether the same problems would arise".

Davidson agrees that "AAV has undergone the most evaluation and seems practical for the delivery of RNAi". But Partridge is more critical. "With AAV and retrovirus vectors you're just rolling the dice," he says, because of the risk of insertional mutagenesis and cancer. Benitec counters that AAV is not a human pathogen, and cannot replicate or spread once inside cells. It rarely integrates into DNA, and when it does, it is at natural chromosomal breakpoints rather than transcriptional 'hotspots' where activation of potential oncogenes is likely to occur. Benitec expects an Investigational New Drug application to test the AAV-based RNAi therapy in humans "to gain approval fairly quickly" because of the previous use of the virus in trials of gene therapy. It hopes to begin clinical trials by 2006.

There is still some way to go before RNAi's full clinical applications are defined. The current scene is unnervingly reminiscent of the excitement more than a decade ago about ribozymes and antisense (see 'A rival to antisense?', page 603). "There's a lot of hype and a lot of hope. So far I'm not convinced that RNAi will be better than antisense," says Klaus Giese, chief scientific officer at atugen in Berlin. "Everyone is waiting for a strong proof of principle *in vivo*." But however the therapies pan out, it is becoming clear that RNAi is already a powerful tool for probing disease pathways. ■

Julie Clayton is a science writer based in Bristol, UK.

Company	Products/activity	Location	URL
RNAi products and services			
Alnylam	Development of siRNA for therapeutic use	Cambridge, Massachusetts	www.alnylam.com
Ambion	Kits, vectors, siRNAs and reagents for RNAi	Austin, Texas	www.ambion.com
atugen	siRNA-based drug discovery; antisense RNAs and siRNAs	Berlin, Germany	www.atugen.com
BD Biosciences	Culture media, radioassay kits, FACS range of flow cytometers; RNAi kits,	San Jose, California	www.bd.com
Benitec	Proprietary gene silencing RNAi technology	St Lucia, Queensland	www.benitec.com.au
Cell Signaling Technology	siRNA kits; antibodies	Beverly, Massachusetts	www.cellsignal.com
Cenix BioScience	Genome-based high-throughput applications of RNA interference; siRNA design	Dresden, Germany	www.cenix-bioscience.com
Combimatrix	Customized DNA microarrays, RNAi	Mukilteo, Washington	www.combimatrix.com
Dharmacon	siRNA kits and arrays; custom RNAs; large-scale RNA synthesis	Lafayette, Colorado	www.dharmacon.com
Eurogentec	Custom siRNA for RNAi; antibodies; custom antibody-based protein arrays	Seraing, Belgium	www.eurogentec.be
Galapagos Genomics	PhenoSelect adenovirus vector/transgene system for target validation and RNAi	Mechelen, Belgium	www.galapagosgenomics.com
Genospectra	QuantiGene Reagent System for gene expression measurements in RNAi, predictive toxicology and drug development	Fremont, California	www.genospectra.com
Imgenex	Plasmid-based siRNA kits; polyclonal and monoclonal antibodies; ELISA kits	San Diego, California	www.imgenex.com
Intradigm	Discovery and development of RNAi-based therapeutics	Rockville, Maryland	www.intradigm.com
Invitrogen	Kits and reagents for RNAi, cloning, genomics, proteomics, molecular and cell biology	Carlsbad, California	www.invitrogen.com
InvivoGen	siRNA cloning vectors, custom siRNA in plasmid and viral vectors	San Diego, California	www.invivogen.com
Mirus Bio	Products for gene transfer and RNAi	Madison, Wisconsin	www.mirusbio.com
MWG Biotech	Custom oligonucleotide synthesis; online sRNA design; Dharmacon siRNA kits	Ebersberg, Germany	www.mwg-biotech.com
NeoPharm	Development of drug delivery via liposomes	Lake Forrest, Illinois	www.neopharm.com
New England Biolabs	Reagents and kits for molecular biology; HiScribe RNAi transcription kit	Beverly, Massachusetts	www.neb.com
Novagen	Reagents and kits for molecular biology; RiboJuice siRNA transfection reagent	Madison, Wisconsin	www.novagen.com
OligoEngine	pSUPER RNAi vector system for gene silencing; custom oligonucleotides	Seattle, Washington	www.oligoengine.com
Open Biosystems	Clones, RNAi, chromosome paints, antibody services	Huntsville, Alabama	www.openbiosystems.com
Prologo	Custom and specialty DNA and RNA oligos; siRNA oligos for RNAi	Boulder, Colorado	www.prologo.com
Promega	Vectors, reagents and kits for genomics, proteomics and cellular analysis; kits for RNAi	Madison, Wisconsin	www.promega.com
QIAGEN	siRNA kits, custom oligos, SNP analysis, molecular biology workstations	Venlo, The Netherlands	www.qiagen.com
RNA-TEC	Custom oligonucleotide synthesis; RNAi for <i>in vivo</i> experiments; RNA-peptides	Leuven, Belgium	www.rna-tec.com
Sima Therapeutics	Development of siRNAs as therapeutics	Boulder, Colorado	www.sima.com
Spring Bioscience	Knock-down kits for producing siRNAs	Fremont, California	www.springbio.com
Stratagene	Tools and reagents for molecular biology, genomics, proteomics, drug discovery and toxicology; GeneEraser siRNA transfection reagents	La Jolla, California	www.stratagene.com
Targeting Systems	Targeted range of transfection reagents	Santee, California	www.targetingsystems.com
VRxSys	Genetic therapy development	Gaithersburg, Maryland	www.virxsys.com
Antisense			
Active Motif	Peptide nucleic acids for gene silencing; kits and reagents for cell fractionation; nucleic acid isolation	Carlsbad, California	www.activemotif.com
Biognostik	Antisense oligonucleotide custom synthesis kits and services	Gottingen, Germany	www.biognostik.com
Cytogenix	Target validation services via proprietary antisense RNA technology	Houston, Texas	www.cytogenix.com
ExpressOn BioSystems	Scalable RNA structure mapping for design of antisense and siRNA	Roslin, UK	www.expresson.co.uk
GeneTools	Morpholino antisense oligonucleotides	Philomath, Oregon	198.88.150.227/genetools
Integrated DNA Technologies	Custom oligonucleotides including phosphorothioate	Coralville, Iowa	www.idtdna.com
Isis Pharmaceuticals	Discovery and development of antisense RNA therapeutics	Carlsbad, California	www.isispharm.com
Santaris Pharma	Discovery and development of antisense LNA (locked nucleic acid) oligonucleotide therapeutics	Hørsholm, Denmark	www.santaris.com
Exiqon	Microarray slides and chips; LNA oligonucleotides	Vedbaek, Denmark	www.exiqon.com
Sigma-Genosys	Tools for molecular biology; phosphorothioate antisense oligonucleotides	The Woodlands, Texas	www.sigma-genosys.com
SomaGenics	Target validation services based on proprietary RNA Lasso antisense technology	Santa Cruz, California	www.somagenics.com
General			
Agilent	2100 Bioanalyzer for RNA, DNA and proteins; instruments for genomics and proteomics	Palo Alto, California	www.agilent.com
Apogent Technologies	Labware and equipment for life sciences	Portsmouth, New Hampshire	www.apogent.com
Applied Biosystems	DNA and protein sequencers; mass spectrometers; chromatography and PCR analysis; peptide and nucleic-acid synthesizers	Foster City California	home.appliedbiosystems.com
ArmaGen Technologies	Drug delivery	Santa Monica, California	www.armagen.com
Asys Hitech	Microplate readers, washers and dispensers	Eugendorf, Austria	www.asyshitech.com
Axon Instruments	ImageExpress high-throughput system for live-cell assays	Foster City, California	www.axon.com
Beckman Coulter	Automated tools for molecular biology; equipment for oligonucleotide synthesis	Fullerton, California	www.beckmancoulter.com
Bibby Sterilin	Laboratory glassware, consumables and equipment for life-sciences research	Stone, UK	www.bibby-sterilin.co.uk
Biolog	Bacterial identification using phenotype microarrays	Hayward, California	www.biolog.com
Bio-Rad	Products, instruments and software for life-sciences research	Hercules, California	www.bio-rad.com
BioTec	Liquid-handling lab instruments, pipetting workstation, dispenser, microplate washer	Tokyo, Japan	www.biotech.co.jp
BMG Labtechnologies	Microplate and array readers and handling systems	Offenburg, Germany	www.bmg-labtechnologies.com
Brinkmann	Laboratory instrument suppliers; software; consumables	Westbury, New York	www.brinkmann.com
BTX	Equipment for electroporation	Holliston, Massachusetts	www.btxonline.com
Corning	Labware and laboratory equipment for life sciences	Corning, New York	www.corning.com
CyBio	Pipetting, liquid handling, incubation and imaging systems for automated screening	Jena, Germany	www.cybio-ag.com
Cyto Pulse Sciences	Equipment for electroporation	Columbia, Maryland	www.cytopulse.com

Company	Products/activity	Location	URL
DiscoverX	Cell-based assay development	Fremont, California	www.discoverx.com
Dynex Technologies	Microplate readers, washers and automated workstations	Chantilly, Virginia	www.dynextechnologies.com
EMD Biosciences	Novabiochem, Calbiochem, Novagen and Oncogene Research products	San Diego, California	splash.emdbiosciences.com
Eppendorf	Laboratory instrumentation and consumables for molecular and cell biology	Hamburg, Germany	www.eppendorf.com
EUGENEX Biotechnologies	Development of test cell lines	Taegerwilen, Switzerland	www.eugenex.com
GE Healthcare Bio-Sciences	Instruments, equipment and products for life-sciences research	Little Chalfont, UK	www.amersham.co.uk
Gilson	Automated liquid-handling and pipetting instruments	Middleton, Wisconsin	www.gilson.com
Greiner Bio-One	Consumables for high-throughput research and diagnostics; microplates	Frickenhausen, Germany	www.gbo.com/bioscience ●
Hamamatsu	Imaging and analysis systems	Hamamatsu City, Japan	www.hamamatsu.com
Hamilton Company	Microlab automated liquid-handling workstations	Reno, Nevada	hamiltoncomp.com
Hybrigenics	<i>In silico</i> prevalidation, and experimental validation services	Paris, France	www.hybrigenics.com ●
Incyte	Annotated gene and expressed sequence tag databases	Palo Alto, California	www.incyte.com
Molecular Devices	Equipment and instruments for microplate-processing and imaging	Sunnyvale, California	www.moleculardevices.com
Nanodrop	ND-1000 UV/visual spectrophotometer	Wilmington, Delaware	www.nanodrop.com
Nalge Nunc International	Labware	Rochester, New York	www.nalgenunc.com
OptiCell	Automated cell-culture devices	Westerville, Ohio	www.opticell.com
Origene	Human cDNA clones; reagents and tools for genomics	Rockville, Maryland	www.origene.com ●
Pierce Chemical	Components and consumables for automated systems in the life sciences	Rockford, Illinois	www.piercenet.com
PerkinElmer Life Sciences	Automated systems for liquid handling, sample preparation and live-cell imaging	Boston, Massachusetts	lifesciences.perkinelmer.com
PreAnalytiX	Kits for nucleic acid isolation from blood samples	Hombrechtikon, Switzerland	www.preanalytix.com ●
Roche Diagnostics	Reagents and kits for molecular biology	Lewes, UK	www.roche-applied-science.com
Robbins Scientific	Labware, instruments and automated liquid-handling and dispensing equipment	Sunnyvale, California	www.matrixtechcorp.com
Stratagene	Tools and reagents for molecular biology, genomics, proteomics, drug discovery	La Jolla, California	www.stratagene.com
Takara Bio	Reagents, kits and custom services for molecular biology	Shiga, Japan	www.takara-bio.co.jp/english
Tecan	Laboratory automation, Cellerity cell-culture system	Männedorf, Switzerland	www.tecan.com
Wako Chemicals	Speciality chemicals, bioproducts and clinical diagnostic reagents	Richmond, Virginia	www.wakousa.com
Wave Biotech	Equipment for cell culture: Wave Bioreactor disposable cell-culture systems	Bridgewater, New Jersey	www.wavebiotech.com
Xantos	High-throughput gene expression assays for drug discovery	Munich, Germany	www.xantos.de ●
Xcellsz	Immortalized cell lines, cell-based assays, screening	Newcastle upon Tyne, UK	www.xcellsz.com

● see advertisement

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejp.com

naturejobs

Two-way traffic

It used to be that the career path between science and industry was a one-way street. If you left academia for the corporate research bench, you simply couldn't get back. This is now no longer true, and many industrial scientists have found a way to return to their academic roots (see *Nature* 430, 706–707; 2004). But what of those who quit not only academia, but also the lab? The times, it seems, are changing for them as well.

Take scientific publishing. There are, I am sure, one or two disgruntled authors, suffering rejection of their work, who might claim that journal editors would no longer be able to cut it in the lab. But in fact some editors continue to conduct research, independently of their publishing jobs. And a few even tread the path back to full-time research.

The rule seems to apply to scientists at many different career stages. For example, an intern who recently completed a stint as an editor on *Nature Structural & Molecular Biology* has accepted a post in a lab, where she hopes to put her freshly honed communication skills to good use (see page 612). She is not alone. A quick, informal survey of people who have left *Nature* in recent years reveals that, as might be expected, many moved on to jobs at other journals, magazines, newspapers or even on television. But a surprisingly high number took positions at leading labs and prominent institutions. The posts range from postdoc to senior administrator. And at least one continues to pursue an active research programme while carrying a full-time editorial load.

Nature is not an anomaly; other scientific journals can tell similar stories. For a researcher who wants to experience something outside the lab environment, scientific publishing is a good career option. Whatever stage in your career you are at, the job market's increasing flexibility should allow you to travel in any direction.

Paul Smaglik
Naturejobs editor



Contents

CAREER VIEW

Industry & Interns

The intern experience

Graduate Journal

Changing directions

Movers

Wolfgang Heckl

p612

WWW.NATUREJOBS.COM

Career centre

Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

GRADUATE JOURNAL

Changing directions

I recently made the tough decision to abandon the research project I started two years ago. The interdisciplinary nature of the biophysical problem I was investigating initially appealed to me. I knew that if I answered key questions, new avenues, both scientifically and professionally, would open. But I also understood from the beginning that working on the frontier meant a high risk of failure.

Until a couple of months ago, the problem remained elusive. Even after that, the data we were generating only showed a few hints of a phenomenon, not definite evidence something was happening.

So, finally, I was confronted with the decision to continue or to switch to another project. Even though I am quite tenacious, I made this decision fast. A new project means the opportunity to dive into a subject in which more knowledge and experience is available.

I have friends who were reluctant to change in a similar situation. Yet I am not unhappy. After all, I learned a lot, especially how to seek out and talk to researchers in other disciplines and assemble pieces of information into a whole. A new project offers new challenges and opportunities. I like a white piece of paper that waits to be filled with new concepts.

Philipp Angerer is a third-year PhD student in biotechnology at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland.

INDUSTRY & INTERNS

The intern experience

Cell biologist Jon Reynolds preferred reading and writing about science to doing research in the lab. Cancer biologist Nadia Cervoni chose to pursue an interest in science communication. Neuroscientist Stacie Grossman wanted to enter the publishing arena. And I sought a broader exposure to science and wanted to hone my writing skills before starting an independent research position.

Although our scientific backgrounds and our motives for applying for a *Nature* editorial internship were quite diverse, we had one thing in common. With due supervision, we did the same work as the full-time editors of these journals, shepherding manuscripts through the decision-making process, writing press releases and research highlights, commissioning and editing

review articles and corresponding directly with referees and authors.

Jon's six-month spell at *Nature Cell Biology* gave him insights into both the editorial and production aspects of a leading scientific journal, and he subsequently found his niche in production. Nadia worked at *Nature Biotechnology*, where she has now secured a full-time post as an assistant editor. Stacie covered neuroscience during her internship at *Nature Medicine* and, in her current role as associate editor, she has broadened her expertise to include cardiovascular biology, autoimmunity and metabolic disease.

The experiences of the other interns helped to prepare them for their current roles in publishing. My internship at *Nature Structural & Molecular Biology* will help me when I return to the lab.

The ability to write a paper clearly and logically,

to choose the most appropriate journal, and to respond effectively to the comments of referees are vital skills for a group leader — and skills that the internship showed me from the other side of the peer-review process.

As an intern, I was challenged to think analytically about a wider range of scientific topics and approaches than I was ever exposed to as a graduate student or postdoc. Feedback from expert referees and experienced editorial colleagues confirmed whether or not my manuscript assessments were valid.

I hope to turn my behind-the-scenes editorial experience into successful refereeing and publishing experiences in the future.

Rosemary Clyne was an editorial intern at *Nature Structural & Molecular Biology*. She is now a research fellow at the Conway Institute in Dublin, Ireland.

MOVERS Wolfgang Heckl, director-general, Deutsches Museum, Munich, Germany



Wolfgang Heckl's boyhood obsession with taking apart household items like radio sets drove his parents to despair. But this curiosity proved to be the seed of a successful scientific career.

Although he turned to dissecting the most powerful and advanced microscopic tools available, this 46-year-old nanoscientist has preserved his early love for screwdrivers and soldering irons. Over the years, he has become the proud owner of a collection

of antique radio sets and classic 1950s jukeboxes, all lovingly repaired, maintained and displayed in his Munich home.

Heckl's scientific tinkering was nurtured by his mentor Gerd Binnig, who in 1986 won a share of the Physics Nobel Prize for developing scanning tunnel microscopy. The instrumentation allows its users to see and manipulate individual atoms and molecules.

Heckl became an early adopter of Binnig's technique, and of his pioneering spirit, when he joined Binnig's group at IBM Research in Munich as a postdoc in 1989. "I owe Binnig a lot," Heckl says. "He really taught us to always make use of our creativity, and encouraged us to dare the unusual."

Binnig's ethos was evident when Heckl created the first visual image of molecules of a DNA base, guanine. Heckl's microscopy-based studies, conducted with New Zealand biochemist

Stephen Sowerby on the spontaneous self-assembly of molecules on a crystal lattice, have provided experimental insight into the possible origin of life. Some scientists believe that well-structured mineral layers, like nanoscale egg boxes, may have provided the 'template' for the first formation of DNA.

As director-general of Deutsches Museum, Germany's largest scientific museum, Heckl is returning to his roots: the museum displays a scanning tunnelling microscope like the one Binnig used in the early 1980s.

Heckl plans for modern science to be better represented. But the museum's invaluable historic exhibits will remain its main attraction. Heckl can continue to indulge his interest in what he now calls 'techniquities' — a mix of technology and antiquities such as the experimental equipment used to discover fission in 1938. But now he needn't worry about parental repercussions.

CV 1993–: Full professor of experimental physics, Ludwig Maximilian University of Munich, Crystallography and Applied Mineralogy section

1990–93: Assistant, Ludwig Maximilian University of Munich, Department of Physics

1989–90: Postdoctoral researcher, IBM Research, Munich

1988–89: Postdoctoral researcher, University of Toronto, Department of Chemistry